Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 15 July 2003
Accepted 11 September 2003

Ab initio protein phasing at 1.4 Å resolution: the new phasing approach of *SIR2003-N*

Maria C. Burla,[a] Benedetta Carrozzini,[b] Rocco Caliandro,[b] Giovanni L. Cascarano,[b] Liberato De Caro,[b] Carmelo Giacovazzo[c,b]* and Giampiero Polidori[a]

[a]Dipartimento di Scienze della Terra, Piazza Università, 06100 Perugia, Italy, [b]Istituto di Cristallografia, CNR, c/o Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and [c]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy. Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

New algorithms for solving *ab initio* protein crystal structures have been identified and implemented in a modified version of the program *SIR2002*. They succeed in solving numerous protein structures diffracting at atomic resolution; the solution was also attained when data were cut at 1.4 Å resolution. The direct-space refinement procedure of *SIR2003-N* takes advantage of using the envelope of the protein, calculated during the phasing process from the current phases. The electron-density map is modified by assuming different weights for pixels within the envelope or out of it, so tentatively depleting the intensities of the false peaks. The map is then inverted and the resulting phase sets may improve their values. The new phasing strategy is also based on an optimal use of some figures of merit, one of which may be successfully applied in the early stages of the phasing process: only the most promising trials are submitted to the complete phasing procedure, so saving computing time. *SIR2003-N* has been successfully applied also in solving some protein structures diffracting at 1.4–1.5 Å resolution.

1. Notation

$F_{\mathbf{h}}$: structure factor with vectorial index $\mathbf{h} = (hkl)$;
$E_{\mathbf{h}} = R_{\mathbf{h}}\exp(i\phi_{\mathbf{h}})$: normalized structure factor;
Res = data resolution;
$I_i(x)$: modified Bessel function of order i;
$D_1(x) = I_1(x)/I_0(x)$.

2. Introduction

The ability of direct methods (DM) to solve *ab initio* protein crystal structures is widely proved, provided data resolution is not worse than 1.2 Å and the structure complexity does not exceed 2500 atoms in the asymmetric unit. Four computer programs [*SnB* (Weeks *et al.*, 1994; Rappleye *et al.*, 2002), *SHELX-D* (Sheldrick, 1998), *ACORN* (Foadi *et al.*, 2000), and *SIR2002* (Burla *et al.*, 2002; Burla, Camalli *et al.*, 2003)] have supplied documentary evidence of this ability. The condition on the data resolution restricts the impact of DM on macromolecular crystallography more than the limit on the structural complexity: indeed, only a small percentage of protein structures is able to produce measurable diffraction data at atomic resolution.

The resolution limit of 1.2 Å is commonly considered a necessary condition for the success of DM in the crystal structure determination. In 1990, Sheldrick formulated a precise operational condition according to which:

'... if fewer than half the number of theoretically measurable reflections in the range 1.1 to 1.2 Å are observed [*i.e.* have $F > 4\sigma(F)$], it is very unlikely that the structure can be solved by direct methods. This critical ratio may be reduced somewhat for centrosymmetric structures, and for structures containing heavier atoms'.

This rule has been recently reconsidered by Morris & Bricogne (2003), who provide its structural basis in terms of typical bonding distances and of interferences produced by distance beats that occur in the radial pair distribution functions in proteins and typical organic molecules. Nevertheless, some pioneering papers by Mukherjee & Woolfson (1995) and Mukherjee *et al.* (1999, 2000) have explored the possibility of overcoming the stated resolution limits of DM. Of particular concern is the paper by Usón *et al.* (1999), where the crystal structure solution of Hirustasin is obtained *via* data collected up to 1.4 Å resolution at room temperature. A more recent paper by Burla, Carrozzini, Cascarano *et al.* (2003) has explored the possibility of systematically overcoming the stated resolution limits of DM: it describes the modules and the algorithms of a new program, named *SIR2003* (an evolution of *SIR2002*), and documents its ability to solve ten protein structures by cutting off their data at 1.4 Å resolution: the complete set of experimental data varied from 0.83 to 1.20 Å.

That pilot study proved that the algorithms seem capable of overcoming the atomic resolution limit, but it still remained to prove their efficiency on crystals diffracting below 1.20 Å. The first tests by *SIR2003* on such structures, quoted in this paper, are disappointing: the structures were resistant to any default attempt. The reason for the failure may be identified in the poorer quality of the data relative to proteins diffracting at non-atomic resolution and on the incompleteness of such measurements, apparently confirming Sheldrick's rule. We were then obliged to change once more our algorithms to face the new difficulties: this paper is a report of the new approach (implemented into the version *SIR2003-N*) and of the experimental results we obtain. The new algorithms presented in the paper have been successfully applied in solving some protein structures diffracting at 1.4–1.5 Å resolution, so violating the above-mentioned empirical rule.

In §3, we will briefly recall the *SIR2003* approach and the algorithms there implemented, together with those employed by the new version, *SIR2003-N*. The applications will be described in §4. Finally, in §5, the conclusions appear.

3. The *SIR2003-N* approach

As for *SIR2002* and *SIR2003*, the program *SIR2003-N* also uses different phasing procedures for solving structures of different sizes (from small molecules to macromolecules). In this paper, we will focus our attention only on proteins. In order to better describe the main modules of the new program, it is useful to briefly recall the characteristics of *SIR2003*.

3.1. *SIR2003*

For solving protein crystal structures, *SIR2003* uses the following modules:

(*a*) TT: a triple tangent formula to produce useful sets of phases, starting from random sets;

(*b*) EDM: an electron-density modification procedure to refine and extend phases. The new phases are calculated by inverting both positive and negative electron-density regions of the unit cell. Powering of the electron density is also performed (Refaat & Woolfson, 1993);

(*c*) HAFR: to express a selected number of large-intensity electron-density peaks (*i.e.* $N_{\rm asym}/6$ peaks, where $N_{\rm asym}$ is the number of non-H atoms in the asymmetric unit) in terms of the heaviest atomic species and suitable occupancy factors;

(*d*) LSQH: to refine the isotropic displacement parameters of the heavy atoms *via* a least-squares procedure.

The application of the modules (*b*), (*c*), (*d*) is repeated a fixed number of times after the triple tangent application; the relative procedure is here called DSR (*direct space refinement*). After the application of DSR, the goodness of a trial solution is assessed by the RAT figure of merit (if RAT is larger than a suitable threshold the program stops):

$$\mathrm{RAT} = \mathrm{CC}/\langle R_{\rm cal}^2\rangle, \qquad (1)$$

where

$$\mathrm{CC} = \frac{(\langle R_{\rm obs}^2 w^2\rangle - \langle R_{\rm obs}^2\rangle\langle w^2\rangle)}{(\langle R_{\rm obs}^4\rangle - \langle R_{\rm obs}^2\rangle^2)^{1/2}(\langle w^4\rangle - \langle w^2\rangle^2)^{1/2}} \qquad (2)$$

is the correlation coefficient between the $R_{\rm obs}$'s in the interval (0.3, 1.2) and the corresponding Sim-like coefficients (Sim, 1959) $w = D_1(2R_{\rm obs}R_{\rm cal})$. $R_{\rm cal} = |E_{\rm cal}|$ are the moduli of the normalized structure factors, available after the last EDM cycle. The average at the denominator of (1) is made on 30% of reflections with smallest $|F_{\rm obs}|$. It is important to stress that these weak reflections are never actively used in the phasing process in both *SIR2002* and *SIR2003*.

The more promising trial solutions are identified by using the numerator of RAT, *i.e.* by CC. If

$$\mathrm{CC} > \langle \mathrm{CC}\rangle + \sigma_{\rm cc}, \qquad (3)$$

the phasing process (*i.e.* the DSR procedure in *SIR2003*) is iterated for the same trial solution. $\langle$CC$\rangle$ is the average value of CC calculated over the previous *SIR2003* trials and $\sigma_{\rm cc}$ is the corresponding standard deviation. The iteration stops when RAT no longer increases.

3.2. *SIR2003-N*

Several new algorithms are introduced into *SIR2003-N* to improve its efficiency with structures diffracting at a resolution not better than 1.4 Å:

(*a*) As in *SIR2003*, the triplet invariants are evaluated *via* the P_{10} formula (Cascarano *et al.*, 1984), but the TT algorithm is replaced by a single tangent (ST) module (in the default choice). The rationale is the following. If the number of measurable reflections is smaller (because of the low data resolution), the number of strong reflections (to which the tangent formula is applied) will also be smaller. Then the mechanism of the triple tangent [*i.e.* the set of strong reflections is divided into three parts, and starting random phases are given only to the first third; see Burla *et al.* (2000)] is applied, *via* a random starting approach, to a too small set of strong reflections. Moreover, the reliability of triplets at non-atomic resolution is lower than at atomic resolution, implying the necessity of exploring many more trials. Under these conditions, the advantage of the triple with respect to the single tangent is lost. Since the TT is about an order of magnitude more expensive in computing time than the ST, the new procedure may explore thousands of ST trials in the same time as hundreds of TT trials: however, the new strategy requires an early figure of merit able to select the trials potentially useful just after the ST module. Such an early figure of merit (eFOM) is described in item (*d*) below.

(*b*) The molecular envelope of the protein (Wang, 1985; Leslie, 1987) is used as a mask in the density-modification step. The protein volume is calculated through the Mathews (1968) formula, and the envelope is calculated for each trial solution from the current phases. The average radius used for calculating the envelope has been fixed at 6.0 Å; as in the *FLEX* procedure proposed by Giacovazzo & Siliqi (1997), the electron-density map is modified by assigning unit weights to pixels belonging to the envelope and weights equal to 0.50 to pixels out of it. The purpose is the following: the envelope

information should appoint small weights to the false peaks, so increasing the correlation between the current electron-density map and the 'true' map (calculated on the basis of published phases). Such information cannot be used just after the tangent formula, where a few low-resolution reflections are usually phased and where the mean phase error is normally too large. The molecular envelope is thus calculated for the first time after three macrocycles of the EDM module, and then recursively calculated and applied in the following EDM cycles.

It is worthwhile observing that the envelope is calculated from the set of phases of the current trial: why should it be useful as a restraint for the refinement of the same phases? A reasonable answer may be the following: if the mean phase error is still rather high, say 70°, and is uniformly distributed *versus* the resolution, the lower-resolution reflections are less sensitive to this large phase error and can help to distinguish between solvent and protein region.

(*c*) A new multisolution approach is introduced in *SIR2003-N*. In both *SnB* and *SHELX-D*, a figure of merit is applied after a dual space (internal) loop: if it is the best, then the trial structure is submitted to an additional (external) refinement loop (Sheldrick *et al.*, 2001). In *SIR2003*, as well as in *SIR2002* (Burla, Camalli *et al.* 2003), early figures of merit are not applied: a random seed value primes the first random starting set of phases, which is then submitted to tangent refinement and then to direct-space refinement. At the end of the phasing process, the first trial solution is provided; then a second random seed provides a new starting set of phases, to which the entire phasing procedure is applied, giving rise to the second trial solution, and so on, until the correct phase solution is found and recognized among the various trials (according to its crystallographic residual in *SIR2002* and the RAT FOM value in *SIR2003*). In conclusion, the multisolution mechanism is of sequential type: each random starting set of phases is developed until the entire DSR procedure is applied and eventually iterated. For big structures, each trial may require a relevant computing time: thus the global CPU time necessary for solving a protein could become very large (several weeks) if the correct solution is not found among a relatively small number of trials. The lower reliability of triplets at non-atomic resolution can make the sequential approach much more time consuming than at atomic resolution.

The above-described phasing strategy was compulsory in *SIR2002* and *SIR2003* because we were unable to identify a figure of merit able to rank the trials at an early stage of the phasing process. We have identified and used in *SIR2003-N* the eFOM criterion to discard more than 90% of the trials, even when eFOM is applied just after the tangent refinement. Accordingly, we changed the phasing procedure of *SIR2003-N* as follows: n random seeds are used to generate n sets of random starting phases, to which the tangent formula is applied. The resulting n trials are ranked in decreasing order of eFOM and only a small percentage of them (the top ranked n_s/n) are submitted to direct-space refinement and extension. If the correct solution is found, the program stops; otherwise, other supplementary sets (n' trials) of random phases are considered. These new random sets are, once more, phased by the ST module and ranked by eFOM: only n'_s/n' of them are submitted to the phasing process. Again, if the correct solution is found, the program stops, otherwise the cycle is repeated for n'' supplementary trials, and so on, until the solution is found. This phasing strategy allows us to explore numerous seeds without paying so much in terms of computing time: it thus increases the probability of submitting to direct-space methods a good set of phases.

In our experience, at low resolution, the frequency of good sets of phases can range from one out of several thousands to one out of several hundreds. Accordingly, the chosen strategy distributes the trials in batches of n_i units, rather than simultaneously analysing a unique batch of $(n' + n'' + n''' + \ldots)$ trials. As a default, we found it useful to alternate small and large batches of trials (the results presented in this work have been obtained in default mode): we used the sequence $n = 400$, $n' = 3000$, $n'' = 400$, $n''' = 3000$, and so on, and we explored always the same small subset of trials in the DSR procedure: *i.e.* $n_s = n'_s = n''_s = \ldots \approx 100$. In this way, we obtain the following advantages: (i) to lead to a quick solution those structures for which, owing to the good triplet estimates, the ST module frequently provides a favourable structural model; (ii) to solve those structures for which the triplet invariant estimates are so bad that a large number of trials are necessary before a good set of phases can be produced by the ST module, and subsequently submitted to the DSR procedure; (iii) to solve even those protein structures for which the eFOM ranking is relatively inefficient.

(*d*) The eFOM used in *SIR2003-N* is defined as follows:

$$\mathrm{eFOM} = \frac{\langle R^2_{\mathrm{cal}} \rangle_{\mathrm{strong}}}{\langle R^2_{\mathrm{cal}} \rangle_{\mathrm{weak}}}, \qquad (4)$$

where: (i) R^2_{cal} is the structure-factor modulus calculated by inversion of a small percentage (about 3.5%) of the E map obtainable after the application of the ST module; (ii) the index '*strong*' indicates that the average is calculated over the reflections with the largest $R_{\mathbf{h}}$ modulus (about 70% of the total number of measured reflections); (iii) the index '*weak*' indicates that the average is calculated on the remaining weakest reflections.

The experimental applications described in §4 indicate that eFOM is, as an early figure of merit, more robust than (even if correlated with) CC and RAT. In particular, eFOM is independent of the weights w, which are not very meaningful at an early stage of the phasing process. Conversely, RAT (and CC) is preferable in the last steps of the phasing process, when the weights w are more reliable.

(*e*) *SIR2003-N* stops when the figure of merit ΔRAT (applied at the end of the DSR procedure) identifies the correct solution among the different trials, where

$$\Delta\mathrm{RAT} = \frac{\mathrm{RAT} - \langle \mathrm{RAT} \rangle}{\sigma_{\mathrm{RAT}}}. \qquad (5)$$

Here, RAT is the value for the current trial, ⟨RAT⟩ is the average value of RAT calculated for all the trials explored by

Table 1
Code name and crystallochemical data for protein test structures.

PDB is the file code in the Protein Data Bank, when available; Res is the experimental data resolution in Å; N_{asym} is the number of non-H atoms in the asymmetric unit, H_2O is the corresponding number of water molecules. In the last column, the heavy-atom species are specified.

Structure code	PDB code	Res (Å)	Space group	Residues	N_{asym}–H_2O	Heavy species
Actinomycin[(a)]	1a7z	0.95	$P2_12_12_1$	22	306	Cl_2
App[(b)]	–	0.99	$C2$	36	302	Zn
Aquo[(c)]	1a6k	1.10	$P2_1$	151	1229–191	S_4 Fe
Aspar[(d)]	1fy2	1.20	$C2$	229	1685–192	S_4 Cd
Calmodulin[(e)]	1exr	1.00	$P1$	148	1150–178	S_8 Ca_5
Collagen[(f)]	2knt	1.20	$P2_1$	58	465–50	P S_6
Conotoxin[(g)]	1a0m	1.09	$I4$	34	255–42	S_{10}
Crambin[(h)]	–	0.83	$P2_1$	48	329	S_6
Cutinase[(i)]	1cex	1.00	$P2_1$	214	1441–264	S_5
Cyto553[(j)]	1c75	0.97	$P2_12_12_1$	71	528–125	S_3 Fe
Dorota[(k)]	1ick	0.94	$P2_12_12_1$	12	259	P_{10} Mg
Ferredoxin[(l)]	2fdn	0.94	$P4_32_12$	55	373–94	S_{16} Fe_8
Gav76[(m)]	1i76	1.20	$P2_12_12_1$	136	1315–271	S_3 Ca_2 Zn_2
Gramicidin[(n)]	–	0.86	$P2_12_12_1$	36	317	–
H42q[(o)]	1b0y	0.93	$P2_12_12_1$	85	594–206	S_9 Fe_4
Hipip[(o)]	1cku	1.20	$P2_12_12_1$	170	1229–334	S_{18} Fe_8
Hirustasin[(p)]	1bx7	1.20	$P4_32_12$	55	365–52	S_{11}
Isd[(q)]	1dy5	0.87	$P2_1$	248	1910–374	S_{25}
Jod[(r)]	–	1.15	$C222_1$	∼60	629	I_{17}
Lactalbumin[(s)]	1b9o	1.15	$P2_12_12_1$	123	935–164	S_{10} Ca
Lysozyme[(t)]	–	0.85	$P1$	129	1001–108	S_{10}
Myoglobin[(c)]	1a6m	1.00	$P2_1$	151	1241–186	S_4 Fe
NpII[(u)]	1eb6	1.00	$P2_1$	177	1346–259	S_6 Zn
Oxidoreductase[(v)]	1mfm	1.02	$P2_12_12_1$	153	1106–283	S_2 Cl_2 Cu Zn Cd_4
Parvalbumin[(w)]	2pvb	0.91	$P2_12_12_1$	107	814–211	S_3 Ca_2
Pheromone[(x)]	2erl	1.00	$C2$	40	305–22	S_7
Rnase59[(y)]	1fk3	1.05	$P2_1$	124	950–215	S_{12}
Rubredoxin[(z)]	8rxn	0.91	$P2_1$	52	393–102	S_6 Fe
Toxin II[(aa)]	1aho	1.00	$P2_12_12_1$	64	508–86	S_8
Vancomycin[(bb)]	1aa5	0.90	$P4_32_12$	0	200–55	Cl_8
Vancomycin[(cc)]	1sho	1.09	$P4_32_12$	0	207–108	Cl_8
Erabutoxin[(dd)]	3ebx	1.40	$P2_12_12_1$	62	432–111	S_9
Chole[(ee)]	1lri	1.45	$C222_1$	98	795–46	S_9 Cl
Amicyanin[(ff)]	1aac	1.31	$P2_1$	105	808–138	S_6 Cu
Dnajes[(gg)]	1jes	1.50	$P2_1$	24	486–82	P_{22} Cu_2
Pazur[(hh)]	1paz	1.55	$P6_5$	123	917–93	S_6 Cu

References: (*a*) Schäfer *et al.* (1998); (*b*) Glover *et al.* (1983); (*c*) Vojtechovsky *et al.* (2003); (*d*) Hakansson *et al.* (2000); (*e*) Wilson & Brunger (2000); (*f*) Merigeau *et al.* (1998); (*g*) Hu *et al.* (1998); (*h*) Weeks *et al.* (1995); (*i*) Longhi *et al.* (1997); (*j*) Benini *et al.* (2003); (*k*) Dauter & Adamiak (2001); (*l*) Dauter *et al.* (1997); (*m*) Gavuzzo *et al.* (2000); (*n*) Langs (1988); (*o*) Parisini *et al.* (1999); (*p*) Usón *et al.* (1999); (*q*) Esposito *et al.* (2000); (*r*) Courtesy of O. Nimz; (*s*) Harata *et al.* (1999); (*t*) Deacon *et al.* (1998); (*u*) McAuley *et al.* (2001); (*v*) Ferraroni *et al.* (1999); (*w*) Declercq *et al.* (1999); (*x*) Anderson *et al.* (1996); (*y*) Berisio *et al.* (2002); (*z*) Sheldrick *et al.* (1993); (*aa*) Smith *et al.* (1997); (*bb*) Loll *et al.* (1997); (*cc*) Schäfer *et al.* (1996); (*dd*) Smith *et al.* (1988); (*ee*) Lascombe *et al.* (2002); (*ff*) Durley *et al.* (1993); (*gg*) Atwell *et al.* (2001); (*hh*) Petratos *et al.* (1988).

the program (and not recognized as possible solutions of the protein structure). σ_{RAT} is a function depending (among other parameters) on the resolution (*i.e.*, the higher the resolution, the higher RAT − ⟨RAT⟩). In order to have a figure of merit roughly independent from the data resolution, we modelled it as

$$\sigma_{RAT} = \frac{1}{2.5\mathrm{Res}^2}, \tag{6}$$

where Res is the data resolution in Å.

Thus, a trial is recognized as the correct solution if ΔRAT is sufficiently large. In the ideal case, the correct solutions should show ΔRAT values much larger than those corresponding to the wrong solutions. In other words, the following equation should hold:

$$\Delta\mathrm{RAT} \gg \Delta\mathrm{RAT}_{NS}, \tag{7}$$

where $\Delta\mathrm{RAT}_{NS}$ is the *maximum* value of ΔRAT attained for bad trials. Since ideal cases seldom occur, a ΔRAT histogram is very helpful to identify the correct solution among several trials (see *Applications*).

Finally, let us note that (5)–(7) cannot be applied to the trials top ranked by eFOM. In this case, the quantity ⟨RAT⟩ of equation (5) is estimated by the RAT average calculated after the ST module (say $\langle\mathrm{RAT}\rangle_{tang}$), where

$$\langle\mathrm{RAT}\rangle_{estimated} = 1.5\langle\mathrm{RAT}\rangle_{tang}. \tag{8}$$

4. Applications

The new algorithms of *SIR2003-N* have been tested both at atomic and at 1.4–1.5 Å resolution. We have selected as test structures 31 protein structures diffracting at atomic resolu-

Table 2
For each test structure diffracting at atomic resolution we show: the order number of the correct solution for *SIR2002*; the corresponding order number for *SIR2003-N* before and after the eFOM ranking (in the first case the figure gives the order of the trial as produced by the random phase approach, in the second case it gives the order as ranked by eFOM; the lower the rank, the greater the efficiency of eFOM); the CPU time in hours needed by *SIR2002* and *SIR2003-N* for solving each structure; the correlation factor (CORR) between the best electron-density map provided by *SIR2002* and *SIR2003-N* and the published map; the ΔRAT value that identifies the solution *versus* the maximum value ΔRAT found for the non-solution trials (ΔRAT_{NS}).

Structure code	Trial *SIR2002*	Trial *SIR2003-N* (non-ranked)	Trial *SIR2003-N* ranked	CPU time *SIR2002* (h)	CPU time *SIR2003-N* (h)	CORR *SIR2002*	CORR *SIR2003-N*	ΔRAT	ΔRAT_{NS}
Actinomycin	38	380	2	2.3	0.4	0.84	0.93	4.5	0.0
App	229	30	1	19.7	0.4	0.62	0.86	4.3	0.0
Aquo	>300	2507	105	–	51.3	–	0.86	6.7	0.3
Aspar†	>300	543	156	–	173.0	–	0.86	9.3	3.9
Calmodulin	>300	309	1	–	5.3	–	0.86	9.0	0.0
Collagen	62	105	14	5.6	0.8	0.77	0.84	5.2	0.9
Conotoxin	17	197	1	2.8	0.3	0.80	0.88	7.1	0.0
Crambin	2	11	5	0.5	0.6	0.80	0.87	5.8	1.0
Cutinase	35	286	37	27.8	22.0	0.84	0.92	7.5	2.0
Cyto553	6	63	1	2.3	0.8	0.83	0.88	9.1	0.0
Dorota	193	157	78	14.5	8.8	0.85	0.95	8.7	0.9
Ferredoxin	18	121	1	7.3	1.2	0.84	0.94	7.7	0.0
Gav76	>300	824	135	–	108.1	–	0.83	6.8	0.7
Gramicidin	30	369	46	2.3	6.1	0.84	0.93	2.6	1.2
H42q	5	259	1	2.2	0.9	0.81	0.91	7.2	0.0
Hipip	4	340	76	5.9	52.8	0.77	0.87	10.1	0.9
Hirustasin	86	3002	172	24.9	54.5	0.83	0.89	8.5	1.3
Isd	146	4	16	209.6	15.5	0.76	0.93	8.0	1.5
Jod	29	210	68	9.2	29.4	0.75	0.89	7.2	1.2
Lactalbumin	62	8	1	22.7	1.2	0.79	0.84	7.9	0.0
Lysozyme	122	203	1	31.5	5.0	0.88	0.88	7.6	0.0
Myoglobin	13	36	63	6.0	21.2	0.83	0.91	12.2	0.6
NpII	96	97	37	57.6	18.6	0.86	0.93	8.6	0.6
Oxidoreductase	59	400	1	50.5	1.6	0.85	0.92	11.4	0.0
Parvalbumin	>300	6687	283	–	143.8	–	0.90	8.0	0.3
Rnase59†	>300	1139	100	–	45.1	–	0.89	10.2	0.7
Pheromone	29	28	5	2.3	0.9	0.82	0.91	8.3	0.8
Rubredoxin	2	15	1	0.5	0.3	0.85	0.92	9.0	0.0
Toxin II	230	230	90	40.6	20.2	0.85	0.93	7.8	0.3
Vancomycin 1aa5	18	214	11	3.7	3.2	0.57	0.93	9.0	0.1
Vancomycin 1sho	49	207	48	7.4	6.3	0.85	0.92	12.0	2.2

† Results obtained by using the envelope module.

tion (between 0.83 and 1.20 Å) and 5 proteins diffracting at non-atomic resolution (from 1.31 to 1.55 Å). We will use the atomic resolution experimental data sets to: (*a*) compare the efficiency of *SIR2003-N* with *SIR2002* performances at atomic resolution; (*b*) check *SIR2003-N* against *SIR2003* performances when the experimental data are cut at 1.4 Å resolution [*SIR2002* is clearly not designed to succeed at this resolution (see Burla, Carrozzini *et al.*, 2003)]. Finally, the five proteins diffracting at non-atomic resolution have been included in the set of test structures to study the efficiency of *SIR2003-N* in this difficult experimental situation. The test structures and their main crystallochemical data are listed in Table 1. For each structure, its code name, the PDB code, the experimental data resolution, the space group, the number of residues, the corresponding non-H atoms in the asymmetric unit, and the heavy-atom species are quoted.

4.1. Tests at atomic resolution

The subset of test structures diffracting at atomic resolution are processed by *SIR2002* and *SIR2003-N*: the complete set of diffraction data is used. In Table 2, we show, for each test structure:

(*a*) The order number of the first trial leading to the correct solution for both *SIR2002* and *SIR2003-N*. To contain the CPU time within reasonable limits, only 300 random trials have been explored by *SIR2002*: the notation '>300' means that no solution has been found among the first 300 trials. Owing to the negligible CPU time required by the ST module in *SIR2003-N*, we required the program to explore up to 6800 trials (*i.e.* two pairs of batches 400 + 3000).

(*b*) The order trial of the correct solution as ranked by eFOM in *SIR2003-N*. The comparison between ranked and non-ranked trial numbers shows the satisfactory efficiency of eFOM.

(*c*) The CPU times needed by *SIR2002* and *SIR2003-N* for solving each structure, expressed in hours (all the numerical tests were performed by using a Xeon-1.7 GHz processor, Linux operating system).

(*d*) The correlation between the electron-density map corresponding to the correct solution provided by *SIR2002* and *SIR2003-N* and the published map (the symbol '–' indicates that no solution was found).

Table 3
Experimental data cut at 1.4 Å for structures diffracting at atomic resolution, and original experimental data for the structure diffracting at resolutions between 1.31 and 1.55 Å.

For each test structure we show: the trial solution for the program *SIR2003-N* when eFOM is not and when it is used (the lower the rank, the greater the efficiency of eFOM); the CPU time in hours needed by *SIR2003-N* for solving each structure; the correlation factor (CORR) between the best electron-density maps provided by *SIR2003* and *SIR2003-N* and the true map at 1.4 Å resolution (in parentheses is the value obtained without the use of the envelope); the ΔRAT value that identifies the solution *versus* its maximum value found for the non-solution trials (ΔRAT_{NS}). The symbol '—' indicates that no solution has been found.

Structure code	*SIR2003-N* (non-ranked)	*SIR2003-N* (ranked)	CPU time (h)	CORR *SIR2003*	CORR *SIR2003-N*	ΔRAT	ΔRAT_{NS}
Actinomycin	–	–	–	–	–	–	–
App	50	1	0.2	0.71	(0.71) 0.77	5.2	0.0
Aquo	–	–	–	–	–	–	–
Aspar	3487	272	502.1	–	(–) 0.78	5.1	0.6
Calmodulin	–	–	–	–	–	–	–
Collagen	153	78	17.2	0.61	(–) 0.70	2.7	0.5
Conotoxin	372	1	0.4	0.64	(0.72) 0.77	3.9	0.0
Crambin	396	26	2.7	0.60	(–) 0.77	3.7	1.2
Cutinase	–	–	–	–	–	–	–
Cyto553	98	69	23.5	–	(0.58) 0.68	2.5	0.8
Dorota	–	–	–	–	–	–	–
Ferredoxin	138	1	0.9	0.86	(0.88) 0.89	11.3	0.0
Gav76	1398	166	197.5	0.56	(0.66) 0.80	5.4	0.6
Gramicidin	–	–	–	–	–	–	–
H42q	38	3	1.6	0.73	(0.74) 0.80	5.1	0.0
Hipip	216	20	27.0	0.81	(0.82) 0.86	9.9	1.8
Hirustasin	28	79	40.7	0.85	(0.85) 0.89	11.3	1.2
Isd	–	–	–	–	–	–	–
Jod	19	43	29.6	0.78	(0.77) 0.77	6.8	1.8
Lactalbumin	1338	158	122.7	0.73	(0.73) 0.75	6.0	1.0
Lysozyme	–	–	–	–	–	–	–
Myoglobin	–	–	–	–	–	–	–
NpII	–	–	–	–	–	–	–
Oxidoreductase	121	24	20.4	0.80	(0.79) 0.84	9.0	0.7
Parvalbumin	–	–	–	–	–	–	–
Rnase59	–	–	–	–	–	–	–
Pheromone	304	9	1.5	0.54	(0.69) 0.72	3.2	0.5
Rubredoxin	11	2	0.3	0.62	(0.67) 0.76	4.8	0.0
Toxin II	11966	1035	495.1	–	(0.74) 0.77	4.0	0.9
Vancomycin 1aa5	216	126	25.9	0.79	(0.82) 0.86	8.6	1.0
Vancomycin 1sho	2436	239	45.9	0.76	(0.78) 0.82	6.6	1.0
Erabutoxin	130	83	32.1	–	(0.61) 0.80	5.4	0.6
Chole	3684	308	285.4	–	(–) 0.75	5.3	0.8
Amicyanin	113	100	40.4	–	(0.80) 0.86	6.3	0.5
Dnajes	2858	177	29.0	–	(–) 0.59	1.4	1.1
Pazurin	10	20	23.9	–	(0.80) 0.86	9.8	0.9

(*e*) The ΔRAT of the solution and ΔRAT_{NS} values. When in the table $\Delta RAT_{NS} = 0.0$ it means that the trial solution of the structure has been ranked as the first one.

By using (6)–(8), from Table 2 it follows that the identification of the solution is always clearly obtained for all structures at atomic resolution, since usually $\Delta RAT \gg \Delta RAT_{NS}$. Averaging all ΔRAT_{NS} of Table 2, we obtain about 1.0, indicating that σ_{RAT} gives an estimation of the RAT maximum deviations from ⟨RAT⟩ for non-solution trials. When the trial solution is ranked as the first one, the individuality of the solution is based essentially on the increment of RAT due to the DSR procedure with respect to the value obtained after the tangent refinement [equation (8)]. In fact, for bad trials usually (8) gives a reliable estimation of ⟨RAT⟩.

We note:

(I) the program *SIR2002* is not able to solve 6 of the 31 test structures in the first 300 trials (all with a number of non-H atoms in the asymmetric unit equal to or larger than 1000). However, *SIR2003-N* for these six structures gives solutions in reasonable CPU times.

(II) *SIR2003-N* is highly competitive. The average CPU time, calculated over the subset of structures solved by both programs, is 22.0 h for *SIR2002*, 10.9 h for *SIR2003-N*.

(III) The value of CORR averaged on all the 31 test structures is for *SIR2003-N* about 0.90, and it is about 10% higher than the average CORR obtained with *SIR2002*, so denoting the higher quality of the electron-density map produced *via SIR2003-N*.

(IV) Two of the test structures (Aspar and Rnase59) have been solved by *SIR2003-N* only by using the envelope module.

4.2. Tests at not-atomic resolution

We cut at 1.4 Å the data of the 31 test structures diffracting at atomic resolution and we used the original experimental data for the five structures with data resolution lower than

1.31 Å. To contain within reasonable limits the CPU time necessary to perform the tests, only a maximum of 600 random trials have been explored by *SIR2003* and a maximum of 13600 trials (*i.e.* four pairs of batches 400 + 3000) by *SIR2003-N*. In Table 3, we show, for each test structure:

(*a*) the order number of the correct solution for the program *SIR2003-N* before and after the eFOM ranking, the comparison between ranked and non-ranked numbers shows that eFOM satisfactorily works also at non-atomic resolution;

(*b*) the CPU time needed by *SIR2003-N* for solving each structure, expressed in hours;

(*c*) the values of CORR for *SIR2003* and *SIR2003-N*;

(*d*) the ΔRAT values for the correct solutions and the corresponding ΔRAT_{NS} obtained for the set of wrong solutions.

We note:

(I) the program *SIR2003-N* solves 24 of the 36 test structures, while *SIR2003* only solves 16 of them. In particular, *SIR2003* is unable to solve any of the five structures diffracting at non-atomic resolution. The number of failures is correlated with the structure complexity and the scattering power of the heavy atoms in the asymmetric unit.

(II) the CPU time necessary for *SIR2003-N* to solve most of the structures is small in most of the cases. The CPU time averaged over all the test structures is 55.8 h.

(III) The value of CORR for *SIR2003-N* averaged over all structures is 0.79, about 10% higher than the average value obtained by *SIR2003*, denoting the higher quality of the electron-density map produced *via SIR2003-N*. Of particular interest is the role of the envelope in the phasing procedure. If we look at the values in parentheses in the column 'CORR *SIR2003-N*', we see that using the envelope has two effects: to bring to solution five test structures that, without envelope, would remain unsolved by *SIR2003-N* and to increase the quality of the final electron-density map.

(IV) Among all the test structures, the only one for which the condition given by (7) is completely unsatisfied is Dnajes: ΔRAT = 1.4 against ΔRAT_{NS} = 1.1. This is also the only solved structure for which the quality of the final electron-density map is of mediocre quality (CORR = 0.59). Actually, *SIR2003-N* is unable to improve further the quality of that map and the pair (ΔRAT, ΔRAT_{NS}) correctly identifies this situation.

However, a clearer insight of the role of the pair (ΔRAT, ΔRAT_{NS}) in the identification of the correct solution may be gained by observing the ΔRAT histogram for all the explored trials, given online by *SIR2003-N*. As examples, in Fig. 1 we report the final histograms obtained for Chole (Fig. 1*a*), Erabutoxin (Fig. 1*b*) and Dnajes (Fig. 1*c*). The correct solution can be clearly identified for all three cases, even for the least favourable case (Dnajes).

5. Conclusions and future work

The new algorithms described in the preceding sections, aimed at solving *ab initio* crystal structures of proteins, succeeded when applied: (i) to numerous protein structures diffracting at atomic resolution; (ii) to the same macromolecules when the experimental data were cut at 1.4 Å resolution; (iii) to a set of protein structures diffracting at 1.3–1.5 Å resolution.

The DSR procedure of *SIR2003-N* takes advantage, among other new tools, of: (i) the estimated envelope of the protein, calculated during the phasing process from the current phases; (ii) the new phasing strategy, based on an optimal use of two figures of merit. One of them (the eFOM) may be successfully applied in the early stages of the phasing process: only the most promising trials are submitted to the complete phasing

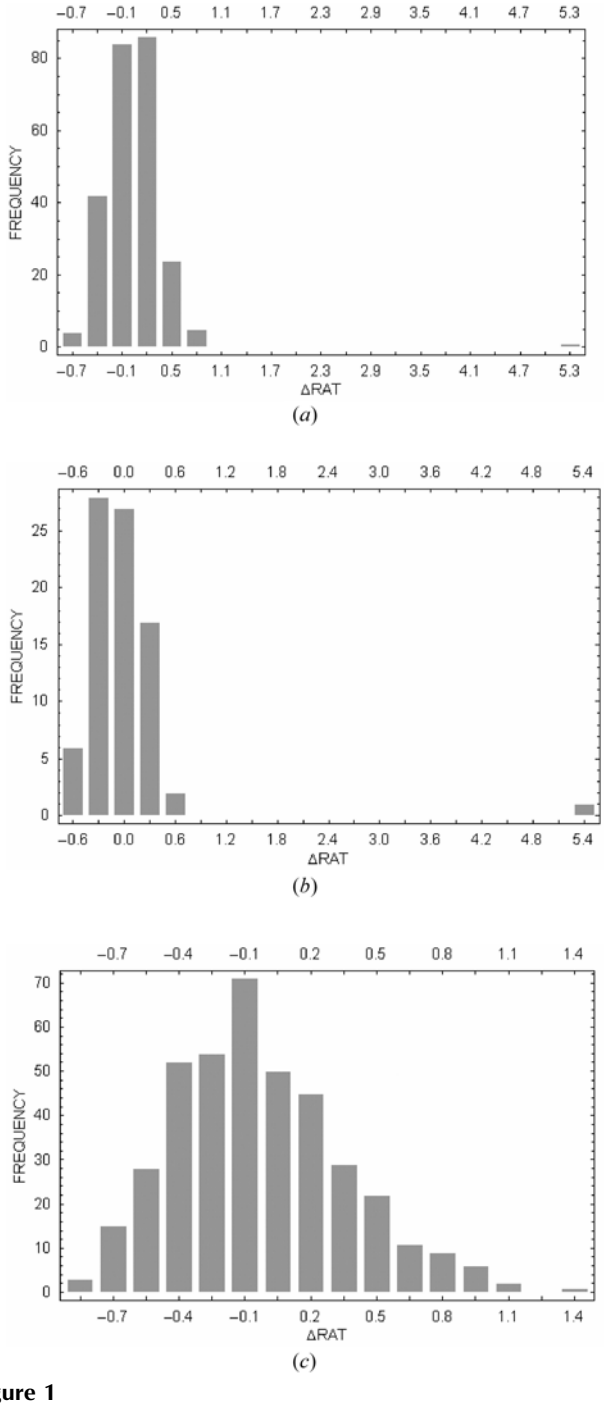

Figure 1
ΔRAT histograms for (*a*) Chole, (*b*) Erabutoxin and (*c*) Dnajes.

procedure, so saving computing time. The second figure of merit (ΔRAT) may be successfully applied to identify the correct solution of the structure.

In spite of some failures, reported in Table 3, we have shown that the *ab initio* crystal structure solution of proteins at non-atomic resolution is a feasible task. This result seems to violate Sheldrick's (1990) rule and its interpretation by Morris & Bricogne (2003). We do not share this conclusion: in our opinion, Sheldrick's rule and its interpretation are probably true if the classical definition of direct methods is assumed, according to which they *directly* derive phases from diffraction modules, without passing through a structural model. Their principal tools were therefore the reciprocal-space relationships, aiming at estimating phase invariants or seminvariants from the prior knowledge of the diffraction moduli, or phases from other phases suitably selected [for a generalization of this approach, see the *neighbourhood principle* by Hauptman (1975) and the *representation theory* by Giacovazzo (1977, 1980)]. If some additional prior information on the molecule is available, classical direct methods try to improve, *via* that information, the probabilistic estimates of the invariants and seminvariants [see Main (1976), Beurskens *et al.* (1976), Camalli *et al.* (1985)].

In more recent years, the reciprocal-space relationships have been integrated with direct-space techniques: *SnB* first considered such an integration as a robust and general technique for solving macromolecular structures, so dramatically generalizing special procedures for phase extension and refinement proposed, for example, by Karle (1970), Sheldrick (1982) and Altomare *et al.* (1991). Nowadays, the usual definition of direct methods tends to include both reciprocal- and direct-space techniques. Since these last can make a simpler but more effective use of some specific structural features (*e.g.* the presence of solvent in the proteins, the form of the protein envelope, ...), modern direct methods can exploit this supplementary source of information, not easily accessible to classical direct methods. In conclusion, we think that Sheldrick's rule, formulated in 1990, holds, for classical direct methods, in a very strict way for equal-atom structures. If a modern definition of DM is assumed, the atomic resolution limit may be violated, particularly when some heavier atoms are present.

Our future work will try to make more robust the present approach for structures diffracting at 1.4 Å resolution, and to succeed also in cases in which the data resolution is lower. In particular, we will try:

(*a*) to identify some new early FOMs to integrate with the eFOM, in order to select with higher efficiency the most promising sets of phases so that we could spend, for the most promising trials, much more CPU time in the DSR module;

(*b*) to integrate direct methods with Patterson techniques (Burla, Carrozzini, Caliandro *et al.*, 2003) in order to overcome the limits of the tangent formula for complex structures at non-atomic resolution;

(*c*) to improve the DSR module, to make easier the phase convergence.

References

Altomare, A., Cascarano, G. L., Giacovazzo, C. & Viterbo, D. (1991). *Acta Cryst.* A**47**, 744–748.

Anderson, D. H., Weiss, M. S. & Eisenberg, D. (1996). *Acta Cryst.* D**52**, 469–480.

Atwell, S., Meggers, E., Spraggon, G. & Schultz, P. G. (2001). *J. Am. Chem. Soc.* **123**, 12364–12367.

Benini, S., Gonzalez, A., Rypniewski, W. R., Wilson, K. S., Van Beeumen, J. J. & Ciurli, S. (2003). In preparation.

Berisio, R., Sica, F., Lamzin, V. S., Wilson, K. S., Zagari, A. & Mazzarella, L. (2002). *Acta Cryst.* D**58**, 441–450.

Beurskens, P. T., van der Hark, Th. E. M. & Beurskens, G. (1976). *Acta Cryst.* A**32**, 821–822.

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Spagna, R. (2000). *Acta Cryst.* A**56**, 451–457.

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Spagna, R. (2003). *J. Appl. Cryst.* **36**, 1103.

Burla, M. C., Carrozzini, B., Caliandro, R., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Polidori, G. (2003). *J. Appl. Cryst.* Submitted.

Burla, M. C., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* A**59**, 245–249.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2002). *Z. Kristallogr.* **217**, 629–635.

Camalli, M., Giacovazzo, C. & Spagna, R. (1985). *Acta Cryst.* A**41**, 605–613.

Cascarano, G. L., Giacovazzo, C., Camalli, M., Spagna, R., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *Acta Cryst.* A**40**, 278–283.

Dauter, Z. & Adamiak, D. A. (2001). *Acta Cryst.* D**57**, 990–995.

Dauter, Z., Wilson, K. S., Sieker, L. C., Meyer, J. & Moulis, J. M. (1997). *Biochemistry*, **36**, 16065–16073.

Deacon, A. M., Weeks, C. M., Miller, R. & Ealick, S. E. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.

Declercq, J. P., Evrard, C., Lamzin, V. & Parello, J. (1999). *Protein Sci.* **8**, 2194–2204.

Durley, R., Chen, L., Lim, L. W., Mathews, F. S., Davidson, V. L. (1993). *Protein Sci.* **2**, 739–752.

Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G., Zagari, A. & Mazzarella, L. (2000). *J. Mol. Biol.* **297**, 713–732.

Ferraroni, M., Rypniewski, W., Wilson, K. S., Viezzoli, M. S., Banci, L., Bertini, I. & Mangani, M. (1999). *J. Mol. Biol.* **288**, 413–426.

Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Yao, J.-X., & Zheng, C.-D. (2000). *Acta Cryst.* D**56**, 1137–1147.

Gavuzzo, E., Pochetti, G., Mazza, F., Gallina, C., Gorini, B., D'Alessio, S., Pieper, M., Tschesche, H., Tucker, P. A. (2000). *J. Med. Chem.* **43**, 3377–3385.

Giacovazzo, C. (1977). *Acta Cryst.* A**33**, 933–944.

Giacovazzo, C. (1980). *Acta Cryst.* A**36**, 362–372.

Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789–798.

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

Hakansson, K., Wang, A. H.-J., Miller, C. G. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 14097–14102.

Harata, K., Abe, Y. & Muraki, M. (1999). *J. Mol. Biol.* **287**, 347–358.

Hauptman, H. (1975). *Acta Cryst.* A**31**, 680–687.

Hu, S. H., Loughnan, M., Miller, R., Weeks, C. M., Blessing, R. H., Alewood, P. F., Lewis, R. J. & Martin, J. L. (1998). *Biochemistry*, **37**, 11425–11433.

Karle, J. (1970). *Crystallographic Computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 13–77. Copenhagen: Munksgaard.

Langs, D. A. (1988). *Science*, **241**, 188–191.

Lascombe, M.-B., Ponchet, M., Venard, P., Milat, M.-L., Blein, J.-P. & Prange, T. (2002). *Acta Cryst.* D**58**, 1442–1447.

Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 134–136.

Loll, P. J., Bevivino, A. E., Korty, B. D. & Axelsen, P. H. (1997). *J. Am. Chem. Soc.* **119**, 1516–1522.

Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A. & Cambillau, C. (1997). *J. Mol. Biol.* **268**, 779–799.
McAuley, K. E., Yao, J.-X., Dodson, E. J., Lehmbeck, J., Østergaard, P. R. & Wilson, K. S. (2001). *Acta Cryst.* D**57**, 1571–1578.
Main, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. Ahmed, pp. 97–105. Copenhagen: Munksgaard.
Mathews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Merigeau, K., Arnoux, B., Norris, K., Norris, F. & Ducruix, A. (1998). *Acta Cryst.* D**54**, 306–312.
Morris, R. J. & Bricogne, G. (2003). *Acta Cryst.* D**59**, 615–617.
Mukherjee, M., Ghosh, S. & Woolfson, M. M. (1999). *Acta Cryst.* D**55**, 168–172.
Mukherjee, M., Maiti, S. & Woolfson, M. M. (2000). *Acta Cryst.* D**56**, 1132–1136.
Mukherjee, M. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 626–628.
Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.
Petratos, K., Dauter, Z. & Wilson, K. S. (1988). *Acta Cryst.* B**44**, 628–636.
Rappleye, J., Innus, M., Weeks, C. M. & Miller, R. (2002). *J. Appl. Cryst.* **35**, 374–376.
Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.
Schäfer, M., Schneider, T. R. & Sheldrick, G. M. (1996). *Structure*, **4**, 1509–1515.
Schäfer, M., Sheldrick, G. M., Bahner, I. & Lackner, H. (1998). *Angew. Chem. Int. Ed. Engl.* **3**7, 2381–2384.
Sheldrick, G. M. (1982). *Computational Crystallography*, edited by D. Sayre, pp. 506–514. Oxford: Clarendon Press.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Sheldrick, G. M. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2001). *International Tables for Crystallography*, Vol. F, pp. 333–345. Dordrecht: Kluwer.
Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.
Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1997). *Acta Cryst.* D**53**, 551–557.
Smith, J. L., Corfield, P. W. R., Hendrickson, W. A. & Low, B. W. (1988). *Acta Cryst.* A**44**, 357–368.
Usón, I., Sheldrick, G. M., de La Fortelle, E., Bricogne, G., Di Marco, S., Priestle, J. P., Grutter, M. G. & Mittl, P. R. (1999). *Struct. Fold. Des.* **7**, 55–63.
Vojtechovsky, J., Berendzen, J., Chu, K., Schlichting, I. & Sweet, R. M. (2003). In preparation.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
Weeks, C. M., De Titta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Weeks, C. M., Hauptman, H. A., Smith, G. D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.
Wilson, M. A. & Brunger, A. T. (2000). *J. Mol. Biol.* **301**, 1237–1256.