Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 11 February 2000
Accepted 8 May 2000

SIR2000, a program for the automatic *ab initio* crystal structure solution of proteins

Maria Cristina Burla,[b] Mercedes Camalli,[c] Benedetta Carrozzini,[d] Giovanni Luca Cascarano,[d] Carmelo Giacovazzo,[a,d]* Giampiero Polidori[b] and Riccardo Spagna[c]

[a]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, [b]Dipartimento di Scienze della Terra, Piazza Università, 06100 Perugia, Italy, [c]Istituto di Strutturistica Chimica 'G. Giacomello', CNR, CP 10 Monterotondo Stazione, 00016 Roma, Italy, and [d]IRMEC c/o Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy. Correspondence e-mail: c.giacovazzo@area.ba.cnr.it

A new phasing procedure is described working both in direct and in reciprocal space. The procedure has been implemented into the program *SIR2000*, the heir to *SIR99*, and it is able routinely to solve *ab initio* crystal structures of proteins without any use of prior information and any user intervention. The moduli and the flow diagram of *SIR2000* are also described and its efficiency tested on several protein diffraction data sets. Success has been attained for crystal structures with up to almost 2000 non-hydrogen atoms in the asymmetric unit and resolution higher than 1.2 Å. The phasing process is analysed to provide a better insight into the role of the various steps of the procedure.

1. Introduction

The recent tremendous increase in computing speed addressed direct-methods evolution towards the development of multisolution techniques (Germain & Woolfson, 1968). New powerful algorithms, implemented into the program *Shake-and-Bake* (Weeks *et. al.*, 1994), allowed an impressive widening of the range of structural complexity amenable to phasing by direct methods. Particular attention should be paid to the minimal principle (DeTitta *et al.*, 1994), which considered the phase problem as one of constrained global optimization, and to the refinement procedure, which alternately uses direct and reciprocal space. More precisely, each trial structure is repeatedly and alternately cycled in real and reciprocal space, *via*:

(*a*) a peak-picking protocol operating on the electron-density map which imposes the atomicity constraint;

(*b*) a parameter-shift optimization technique (Bhuiya & Stanley, 1963) aiming at reducing the value of the minimal function (Hauptman, 1991; DeTitta *et al.*, 1994). Advances in the *Shake-and-Bake* algorithms have been recently published (Weeks & Miller, 1999; Hauptman *et al.*, 1999).

An effective variant of *Shake-and-Bake* is the program termed *half-baked* or *SHELX-D* (Sheldrick, 1998), which cyclically alternates tangent refinement (Karle & Hauptman, 1956) in reciprocal space with the peak-list-optimization procedure, proposed by Sheldrick & Gould (1995), in real space.

A third program, named *SIR99*, capable of solving *ab initio* crystal structures of proteins, has been developed more recently (Burla, Camalli *et al.*, 1999). In *SIR99*, as in its predecessor *SIR97* (Altomare *et al.*, 1999), a strategy different from that adopted by *Shake-and-Bake* and *half-baked* is employed: the tangent-refinement section is followed by the real-space refinement, but without alternation. In *SIR99*, however, the real-space section is much more complex than in *SIR97*, since additional tools needed to improve the solution of macromolecular structures have been included.

The largest structure solvable by *SIR99* was toxin II (Smith *et al.*, 1997): attempts at applying the phasing procedure to larger molecules failed. We then decided to modify the *SIR99* approach in order to:

(*a*) widen the range of structural complexity amenable to phasing;

(*b*) reduce the number of trials needed for crystal structure solution [the correct structural model of toxin II was obtained by *SIR99* after 567.30 h of CPU time on a Compaq Personal Workstation 500 au (SPECfp95: 19.5)]. The strategy and results of our efforts, which have been implemented in *SIR2000*, are described in the present article.

2. Structure and flow diagram of *SIR2000*

Most of the routines of *SIR99* (*DATA*, the data input routine; *MENU*, giving access to the graphical interface; the final refinement routines *LSQ*, *HYDROGEN* and *GEOMETRY*) have not been modified and will not be described here. We will only refer to the *SOLVE* modulus, in which the normalization of structure factors, the setting up of invariant relationships, the application of tangent formula, the calculation and modification of electron-density maps, and the preliminary and automatic least-squares refinement of the trial structure are carried out. After the normalization process, the following steps are executed in *SOLVE* by *SIR99*:

Step 1. Triplet invariants, *via* the *P*10 formula (Cascarano *et al.*, 1984), and quartet invariants, *via* the formula derived by Giacovazzo (1976*a*,*b*) and integrated by the procedure proposed by Altomare *et al.* (1995), are estimated.

Step 2. $N_{\rm large}$ reflections (those with the largest $|E|$'s) are selected and phased *via* a double tangent procedure starting from random phases (Baggio *et al.*, 1978).

Step 3. The phase values thus obtained are processed by the following three procedures in sequence:

(*a*) *EDM* (electron-density modification). 15 supercycles, each constituted by 7 microcycles $\rho \rightarrow \{\varphi\} \rightarrow \rho$, are performed, where ρ is the electron-density map and $\{\varphi\}$ is the set of calculated phases. A fraction of ρ ranging from 2.0 to 2.5% is used in each map inversion, the rest is set to zero [see Shiono & Woolfson (1992) for a related electron-density-modification procedure].

(*b*) *HAFR* (heavy-atom reduced real-space Fourier refinement). It consists of 37 cycles $\rho \rightarrow \{\varphi\} \rightarrow \rho$. The heaviest atomic species are associated with the selected peaks and an occupancy factor is given to each peak in order to take into account peak height, site occupancy and chemical connectivity.

(*c*) *DLSQ* (diagonal matrix least squares). Peaks are now labelled in terms of atomic species according to their heights and the chemical content of the unit cell. Diagonal least-squares-refinement cycles and $2F_o - F_c$ map calculations are cyclically performed to refine the structural model.

Step 4. If the crystallographic residual

$$\mathrm{RES} = \textstyle\sum ||F_{\rm obs}| - |F_{\rm calc}|| / \sum |F_{\rm obs}|$$

is larger than 0.25, a new trial is started, otherwise the procedure stops and the graphical interface displays a picture of the trial structure to allow interaction with the user.

The flow diagram of *SIR2000* is shown in Fig. 1. We note that:

(i) As in *SIR99*, triplet invariants are evaluated by the *P*10 formula, but now up to 300000 relationships can be stored, to allow handling of larger molecules (modulus *INVAR* in Fig. 1).

(ii) A triple tangent procedure is used (instead of the double tangent process used in *SIR99*). $N_{\rm large}$ reflections, with the largest $|E|$'s, are given random starting phase values which are treated according to the following three-step procedure:

(*a*) The first $N_{\rm large}/3$ random phases are refined by tangent recycling.

(*b*) Other $N_{\rm large}/3$ reflections with random phase are added and, of the $N_{\rm large}/3$ phases obtained in step (*a*), the $N_{\rm large}/8$ with largest α values are held fixed for the first ten cycles and then relaxed; $N_{\rm large}/3$ phases are obtained.

(*c*) The last $N_{\rm large}/3$ reflections with random phase are added and $N_{\rm large}/4$ reflections, obtained with largest α values in step (*b*), are held fixed for the first ten cycles, and then relaxed; all $N_{\rm large}$ phase are thus obtained.

(iii) The observed $|E|$ values of the low-resolution reflections are modified (modulus MODE). This action is undertaken because large regions of the unit cell of macromolecular crystals are filled by solvent. Since the molecular envelope is usually unknown at this stage, low-resolution diffraction intensities contain an unpredictable but important solvent contribution.

To restate more reasonable values for protein diffraction intensities by proper subtraction of the solvent effects may be decisive for the success of the phasing process. When the high-resolution structure of the macromolecule is known, information about the solvent structure maybe obtained by assuming

$$F_o = F_p + F_s, \tag{1}$$

where F_o is the observed structure factor, F_p the protein structure factor calculated from the known coordinates and F_s the solvent contribution. Note that equation (1) is a relation between complex quantities and not just between their moduli. The required information about the solvent structure is usually obtained by suitable refinement techniques (see Moews & Kretsinger, 1975).

In our case, the problem is reversed: we try to obtain, for low-resolution reflections, more reasonable values of $|F_p|$ from the $|F_o|$ values, without any prior information on the crystal structure or on the molecular envelope. Several solvent models have been proposed (see Badger, 1997, and literature quoted therein). We use in our procedure the simple solvent model based on Babinet's principle, *i.e.*

$$F_s = -K\exp(-B_s r^{*2})F_p, \tag{2}$$

where K is a suitable (positive) scale factor and $r^* = 2\sin\theta/\lambda$. The exponential term restricts equation (2) to low-resolution reflections and B_s is a displacement parameter accounting for the large motion of the solvent molecules.

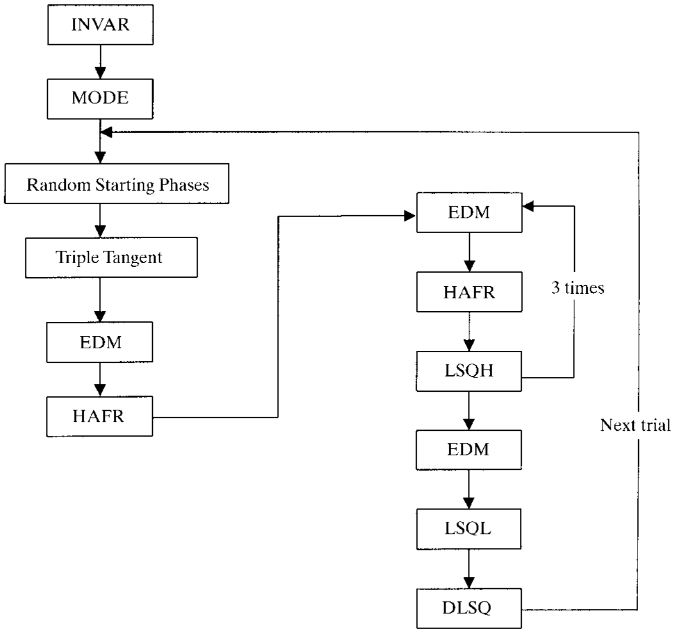

Figure 1
Flow diagram of *SIR2000*.

Then

$$F_o = [1 - K \exp(-B_s r^{*2})]F_p. \quad (3)$$

From now on, $|F_p|$ is the new 'observed' value and will replace $|F_o|$ in all subsequent calculations, provided sensible values of K and B_s are found. Such values may be obtained by requiring, according to Wilson's statistics, that

$$\langle |E_p|^2 \rangle = \langle |E_o|^2 / [1 - K \exp(-B_s r^{*2})]^2 \rangle \cong 1 \quad (4)$$

for any low-resolution 2θ interval. Equation (4), in which the normalized structure factors E have been introduced, suggests that the optimal values of K and B_s can be obtained by imposing that

$$\sum_{\text{intervals}} [\langle |E_p|^2 \rangle - 1]^2 = \min, \quad (5)$$

where the summation ranges over the fixed low-resolution intervals. We include in equation (5) all the reflections from ∞ to 5 Å resolution.

In accordance with Tronrud (1997), K usually varies in the range 0.7–0.9 and B_s in the range 200–500. Owing to the restricted range of the parameters K and B_s, the minimum of equation (5) can be found by a systematic step search within the allowed intervals.

The expected effects of applying equation (3) to the observed intensities may be described in the following way:

(*a*) Low-resolution reflections have $|F_o|$ moduli too small to influence the phasing process. In practice, very low resolution reflections are excluded from such processes: on the contrary, the corresponding $|F_p|$ values, which are much larger than $|F_o|$'s (*e.g.* 5 times larger at $\sin\theta/\lambda = 0$) can contribute to the phasing process and drive it along different directions.

(*b*) Low-resolution reflections are insensitive to fine structural details, but are essential to define the location and the envelope of the molecule. Accordingly, they play a critical role in defining the region of the electron-density map selected for performing the transformation $\rho \rightarrow \{\varphi\}$ in the moduli *EDM* and *HAFR*.

(iv) As in *SIR99*, the real-space-refinement section of *SIR2000* does not alternate with but follows the reciprocal-space refinement. Several relevant differences however should be underlined:

(*a*) *SIR2000* retains the moduli *EDM*, *HAFR* and *DLSQ*, but also contains two new moduli *LSQH* and *LSQL*.

(*b*) While in *SIR99* the sequence of the moduli was very simple (*i.e.* *EDM* → *HAFR* → *DLSQ*), the *SIR2000* sequence is much more complex (see Fig. 1).

(*c*) In *SIR99*, *EDM* consisted of 15 supercycles, each constituted by 7 microcycles $\rho \rightarrow \{\varphi\} \rightarrow \rho$; in *SIR2000*, the number of supercycles is increased to 25. On the contrary, while *HAFR* in *SIR99* consisted of 37 cycles $\rho \rightarrow \{\varphi\} \rightarrow \rho$, in *SIR2000* it has been reduced to only 15 cycles. The practical reason for these changes was that the phasing process proved to be more effective by alternating *EDM* and *HAFR* moduli rather than using them in sequence.

(*d*) The modulus *LSQH* is essentially a least-squares routine which is only applied when heavy atoms (with $Z \geq 11$) are present. While in *HAFR* all the electron-density peaks are associated with the heaviest atomic species, in *LSQH* only the N_H peaks (N_H is the number of heavy atoms in the asymmetric unit) with largest intensity are labelled as heavy atomic species: the rest of the peaks are associated with lighter atoms. *LSQH* minimizes the residual RES by allowing the isotropic displacement parameters of the heavy atoms to vary: six least-squares cycles are performed for each application of the modulus *LSQH.* The rationale of this new modulus is the following: for each cycle *LSQH* → *EDM* → *HAFR*, the thermal parameters suitably modified by *LSQH* will cooperate to generate better $\{\varphi\}$ values, and therefore better subsequent electron-density maps in *EDM* and *HAFR*. We observed that the introduction of *LSQH* makes more frequent the case in which, at the end of the three cycles of *EDM* → *HAFR* → *LSQH*, the largest peaks in the electron-density map correspond to the heavy-atom positions.

(*e*) In *EDM* and *HAFR*, the *SIR99* strategy is retained. The electron-density maps are calculated by using a number of reflections cyclically increasing with the microcycle order. In *EDM*, the fraction of pixels used for electron-density inversion is 2.5% for the first 5 supercycles and it is variable from 2.5 to 3.5% in the other supercycles; the number of phased structure factors varies from N_{large} to 70% of the measured reflections. In *HAFR*, the number of peaks selected in the Fourier map cyclically varies between 50 and 80% of N_{asym}, where N_{asym} is the estimated number of non-hydrogen atoms in the asymmetric unit.

(*f*) It is not infrequent that the last application of *LSQH* ends with a mean phase error larger than 45°: in this case, the automatic structure refinement by *DLSQ* could fail. We introduced the modulus *LSQL* essentially as a step to prepare a more suitable input for *DLSQ* and increasing its convergence speed. While the heavy-atom displacement parameters were refined by LSQH, all the light atoms still preserve the overall thermal parameter estimated by Wilson's statistics. Since this is unrealistic for most of the light atoms, LSQL divides them in groups, according to the corresponding peak heights. Consequently, atoms corresponding to weak peaks will be given larger thermal factors. This usually leads to improved RES values and to smaller mean phase errors, and facilitates the success of the automatic *DLSQ* refinement. When the data resolution is less than or equal to 1.1 Å, *LSQL* may be considered as the final step of the phasing procedure (see below).

3. Experimental applications

SIR2000 has been applied to the set of 29 crystal structures quoted in Table 1: 24 of them are proteins. The range of structural complexity ranges from 200 to about 2000 non-hydrogen atoms in the asymmetric unit. For each test structure, the following information is given: (i) when protein data have been deposited by the Protein Data Bank, the PDB file code is specified; (ii) R is the data resolution in ångstroms; (iii)

Table 1
Code name, space group and crystallochemical data for test structures with $N_{\rm asym} > 200$.

PDB is the file code in the Protein Data Bank, when available; R is the data resolution in Å; $N_{\rm asym}$ is the number of non-hydrogen atoms in the asymmetric unit, H_2O is the number of water molecules. When heavy atoms are present, their species and their number in the asymmetric unit are specified.

Structure code (reference)	PDB	R (Å)	Space group	$N_{\rm asym}$–H_2O	Heavy atoms
64c (1)	–	1.06	$P1$	440	–
Actino (2)	1a7z	0.95	$P2_12_12_1$	306	Cl_2
Alessia (3)	–	0.82	$P2_12_12_1$	212	–
Alpha1 (4)	1byz	0.90	$P1$	449–30	Cl
App (5)	–	0.99	$C2$	302	Zn
Balhimycin (6)	–	0.81	$P1$	312	Cl_6
Bcd (7)	–	0.89	$P1$	208	–
Collagen (8)	2knt	1.20	$P2_1$	465–50	S_6, P
Conotoxin (9)	1a0m	0.90	$I4$	255–42	S_{10}
Crambin (10)	1cbn	0.83	$P2_1$	329	S_6
Cutinase (11)	1cex	1.00	$P2_1$	1441–264	S_5
Cytochrome-c_6 (12)	1cty	1.10	$R3$	672–151	S_3 Fe
Ferredoxin (13)	2fdn	0.94	$P4_3\,2_1\,2$	373–94	S_{16}, Fe_8
Gramicidin A (14)	1alz	0.86	$P2_12_12_1$	317	–
H42q (15)	1b0y	0.93	$P2_12_12_1$	594–206	S_9, Fe_4
Hipip (16)	1cku	1.20	$P2_12_12_1$	1229–334	S_{18}, Fe_8
Isd (17)	–	0.87	$P2_1$	1910–374	S_{25}
Jod (18)	–	1.15	$C222_1$	629	I_{17}
Lactal (19)	1b9o	1.15	$P2_12_12_1$	935–164	S_{10}, Ca
Lysozyme (20)	–	0.85	$P1$	1001–108	S_{10}
Myoglobin (21)	1a6m	1.00	$P2_1$	1241–186	S_4, Fe
Oxidoreductase (22)	1mfm	1.02	$P2_12_12_1$	1106–283	S_2, Cl_2, Cu, Zn, Cd_9
Pheromone (23)	2erl	1.00	$C2$	305–22	S_7
Profl (24)	–	0.92	$P1$	439	P_8
Rubredoxin (25)	8rxn	0.91	$P2_1$	393–102	S_6, Fe
Toxin II (26)	1aho	1.00	$P2_12_12_1$	508–86	S_8
Vancomycin (27)	1aa5	0.90	$P4_3\,2_1\,2$	255	Cl_8
Vancomycin $P1$ (28)	–	0.97	$P1$	404–103	Cl_8
X116a (29)	–	0.99	$P2_1$	290	Br_{10}

References: (1) Tong *et al.* (1997); (2) Schäfer *et al.* (2000); (3) Bacchi *et al.* (2000); (4) Prive *et al.* (2000); (5) Glover *et al.* (1983); (6) Schäfer *et al.* (1998); (7) Gessler (2000); (8) Merigeau *et al.* (1998); (9) Hu *et al.* (1998); (10) Hope (1988); (11) Longhi *et al.* (1997); (12) Frazao *et al.* (1995); (13) Dauter *et al.* (1997); (14) Langs (1988); (15) Parisini *et al.* (1999); (16) Parisini *et al.* (1999); (17) Esposito *et al.* (2000); (18) Nimz (2000); (19) Harata *et al.* (1999); (20) Deacon *et al.* (1998); (21) Vojtechovsky *et al.* (2000); (22) Ferraroni *et al.* (1999); (23) Anderson *et al.* (1996); (24) Burla, Cascarano *et al.* (1999); (25) Sheldrick *et al.* (1993); (26) Smith *et al.* (1997); (27) Loll *et al.* (1997); (28) Loll *et al.* (1998); (29) Haltiwanger, R.C., SmithKline Beecham Pharmaceuticals, unpublished.

space group; (iv) $N_{\rm asym}$–H_2O are the number of non-hydrogen atoms in the asymmetric unit and the number of water molecules, respectively; (v) when heavy atoms are present, their species and their number in the asymmetric unit are specified.

In Table 2, we show the results of our tests: $N_{\rm large}$ is the number of reflections phased by the tangent procedure, Trsol is the number of the trial at which the correct solution was found, RES is the crystallographic residual for the correct solution, Time is the CPU time (in hours) necessary to obtain the correct structural model. All tests were performed using a Compaq Personal Workstation 500 au (SPECfp95: 19.5) and running *SIR2000* in default mode, without any use of features that are specific to the structure under examination (*e.g.* presence of disulfide bridges, information on the solvent region or on the envelope, heavy-atom positions *etc.*) and without any preliminary use of the Patterson superposition techniques. The typical input file for *SIR2000* is shown in Table 3. The character '%' indicates commands: for each command, directives follow, if needed. A brief analysis of Table 2 suggests that:

(*a*) The correct solutions are often obtained with a limited number of trials: more than 100 trials are needed only for three structures (collagen, cutinase and jod).

(*b*) Most of the structures in Table 2 could not be solved by *SIR99*. For the nine structures reported by Burla, Camalli *et al.* (1999), a comparison may be made between *SIR2000* and *SIR99* performances. The increased effectiveness of *SIR2000* is quite evident: Trsol for *SIR2000* is often much smaller than for *SIR99*, and the total CPU time necessary to solve the nine common structures is 115.2 h for *SIR2000* and 1141.3 h for *SIR99*.

A *post mortem* analysis of the phasing process reveals the specific roles of the new moduli in *SIR2000*. In Fig. 2, we focus our attention on the trials leading to the correct solution of the six test structures with more than 1000 atoms in the asymmetric units (cutinase, hipip, isd, lysozyme, myoglobin and oxidoreductase).

The mean phase error (ER) is monitored at some relevant steps of the phasing procedure: at TRTANG (at the end of the triple tangent process), at HAFR1 (after the first application of *HAFR*), at LSQH3 (at the end of the third cycle $EDM \rightarrow HAFR \rightarrow LSQH$), at LSQL1 (after the application of *LSQL*). For brevity, the behaviour of only four of the six structures will

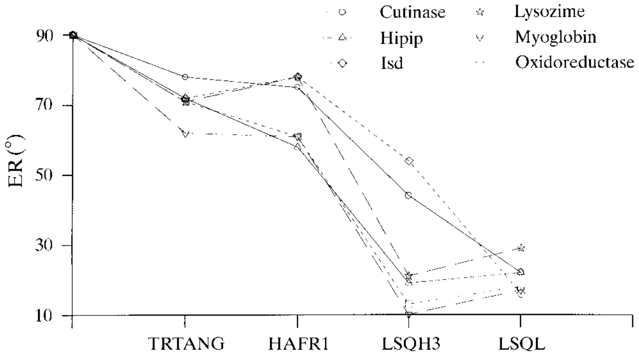

Figure 2
Mean phase error (ER) at some relevant steps of the phasing procedure for six large test structures (see main text).

Table 2
For the test structures in Table 1, we specify for *SIR2000* the order of the trial leading to the correct solution (Trsol), the crystallographic residual (RES) and the time necessary to obtain the solution (in hours); the same information is repeated for *SIR99*, provided a report was made by Burla, Camalli *et al.* (1999).

$N_{\rm large}$ is the number of reflections phased in the triple tangent formula.

		SIR2000			*SIR99*		
Structure code	$N_{\rm large}$	Trsol	RES (%)	Time (h)	Trsol	RES (%)	Time (h)
64c	2180	1	20.17	0.1	–	–	–
Actino	1527	53	14.19	16.1	–	–	–
Alessia	1000	72	15.92	9.1	797	14.51	95.1
Alpha1	2221	4	13.00	0.9	–	–	–
App	1509	87	20.44	30.9	19	19.27	4.0
Balhimycin	1604	1	16.11	0.1	–	–	–
Bcd	1136	4	13.46	0.2	–	–	–
Collagen	2243	112	24.85	26.9	–	–	–
Conotoxin	1298	17	19.94	6.9	–	–	–
Crambin	1633	2	17.70	0.4	190	15.10	40.7
Cutinase	4000	226	19.30	890.3	–	–	–
Cytochrome-c_6	3173	57	27.55	67.7	–	–	–
Ferredoxin	1827	31	17.74	44.4	–	–	–
Gramicidin A	1577	79	20.60	25.3	68	20.59	14.8
H42q	2823	15	22.62	18.8	–	–	–
Hipip	4000	44	20.88	225.6	–	–	–
Isd	4000	61	19.10	374.3	–	–	–
Jod	2901	273	24.98	422.7	–	–	–
Lactal	4000	62	26.40	168.2	–	–	–
Lysozyme	4000	1	19.07	1.0	–	–	–
Myoglobin	4000	25	19.98	54.6	–	–	–
Oxidoreductase	4000	59	25.05	219.7	–	–	–
Pheromone	1525	17	22.39	3.4	–	–	–
Profl	2176	1	18.57	0.1	2	17.48	0.3
Rubredoxin	1919	1	17.09	0.3	15	17.23	4.4
Toxin II	2436	27	18.63	20.5	1024	19.80	567.3
Vancomycin	1299	45	19.49	28.5	909	18.08	414.6
Vancomycin *P*1	2018	4	17.36	0.8	–	–	–
X116a	1455	1	11.44	0.1	1	11.54	0.1

be discussed, the behaviour of the other two being strictly analogous. We notice that:

(i) ER (calculated over the $N_{\rm large}$ reflections given in Table 2) is rather large at TRTANG: ER = 78, 72, 72 and 71° for cutinase, hipip, isd and oxidoreductase, respectively. Any attempt at extending phases by additional tangent cycles failed: the error rapidly grows towards 90°. This behaviour is very frequent also for small structures: *e.g.* at TRTANG ER = 69, 78 and 72° for crambin, gramicidin and vancomycin, respectively, and subsequent phase extension by tangent formula did not succeed. Even if the sole tangent process is unable to solve most of the test structures, its role is of basic importance for *SIR2000*: indeed, the correct solutions are quite frequently found among the trials having the minimum value of ER at TRTANG.

(ii) The phase error at HAFR1 is often larger than at TRTANG. The reader however should not conclude that the first application of the step *EDM* → *HAFR* is of no use: indeed, ER at HAFR1 is calculated over a number of reflections much larger than at TRTANG (over about the number of reflections with $|E| > 0.8$). If we monitor ER for the $N_{\rm large}$ reflections only, we obtain, at HAFR1, ER = 75, 58, 78 and 61 for cutinase, hipip, isd and oxidoreductase, respectively.

(iii) ER (calculated over about 40% of the total unique reflections) is remarkably diminished at LSQH3. Therefore, much of the efficiency of *SIR2000* is due to the three *EDM* → *HAFR* → *LSQH* cycles. However, ER is still quite large for cutinase (60°) and isd (71°): luckily, the subsequent application of the moduli *EDM* and *LSQL* reduces ER to 22 and 16°, respectively (ER now is calculated over about 70% of the total unique reflections).

(iv) If *SIR2000* is run without the use of *MODE* (all the other moduli operating in default), cutinase and hipip cannot be solved in a reasonable time, while isd and oxidoreductase are solved at the same trial order. The application of *SIR2000* to the test structures shows that the efficiency of *SIR2000* is much larger when *MODE* is included; without *MODE*, the solutions, when found, often correspond to higher-order trials.

Let us now clarify the role of some moduli of *SIR2000*. We observe that:

(i) The least-squares procedure in *LSQH* often accelerates the refinement process. Let us consider the case of vancomycin *P*1 as an example. Before the first run of *LSQH*, only 4 correctly located chlorine atoms are found among the 12 largest peaks of the electron-density map provided by *HAFR*; at the end of the third cycle, 7 out of 12 chlorines are among the 12 largest peaks. If step *LSQH* is omitted in the cycles *EDM* → *HAFR* → *LSQH*, then, at the end of the third cycle, the number of chlorine atoms among the largest 12 peaks remains 4 and trial no. 4 ends with a failure.

(ii) The final mean phase error ERF, calculated for all the test structures after the application of *DLSQ*, is usually between 9 and 36°. Five test structures (collagen, cytochrome-c_6, hipip, jod and lactal), all having data resolution lower than or equal to 1.1 Å, converge to larger ERF values, owing to the unfavourable conditions under which *DLSQ* runs. Let us examine in detail the cases of collagen and hipip. Their data resolution is 1.2 Å and the ratios ‘number of observations/number of parameters’ are 6.23 and 7.72, respectively, for isotropic refinements. The application of *DLSQ* in these conditions may be rather critical: refinement could lead to small values of RES without any real structural information. For hipip, *DLSQ* worked well (final values of RES and ER are 0.25 and 25,° respectively). The application of *SIR2000* to collagen data provided a solution only at trial number 112 (see Table 2). At the point *LSQL*, the mean phase error is 47°

Table 3
Typical input file for *SIR2000*.

```
%structure cutinase
%initialize
%job Cutinase
%data
  cell 35.120 67.360 37.050 90.00 93.90 90.00
  spacegroup p 21
  content c 1786 h 3020 n 520 o 563 s 10
  reflections cutinase.hkl
  format(3i4,2f8.2)
%continue
```

Table 4
Some numerical outcomes for test structures with data resolution equal to or lower than 1.1 Å.

Structure code	ERF (°)	ER (°) (*LSQL*)	CORR
Collagen	62	47	0.60
Cytochrome-c_6	59	47	0.80
Hipip	25	22	0.85
Jod	36	28	0.80
Lactal	52	33	0.75

(weighted mean phase error 39°), calculated over 10419 reflections (about 70% of the total unique reflections). The correlation factor CORR between the corresponding electron density and the published one (using all the measured reflections) is 0.6, calculated as

$$\mathrm{CORR} = \frac{(\langle \rho\rho_{\mathrm{mod}}\rangle - \langle\rho\rangle\langle\rho_{\mathrm{mod}}\rangle)}{(\langle\rho^2\rangle - \langle\rho\rangle^2)^{1/2}(\langle\rho^2_{\mathrm{mod}}\rangle - \langle\rho_{\mathrm{mod}}\rangle^2)^{1/2}}.$$

When the refinement by *DLSQ* was completed, the final RES and ER values were 0.25 and 62°, respectively (weighted mean phase error 56°).

A similar situation occurs for cytochrome-c_6, jod and lactal. In Table 4, for each of the five test structures, the values of ERF, ER and CORR after the application of *LSQL* are listed. Table 4 suggests that *DLSQ* must be used with caution when the resolution is less than or equal to 1.1 Å and indicates that an alternative procedure should replace such a modulus for lower-resolution data.

4. Conclusions

A powerful phasing procedure has been described that is able to solve macromolecular structures with up to 2000 non-hydrogen atoms in the asymmetric unit, provided the data resolution is higher than 1.2 Å. The program *SIR2000* proved to be robust since it works on a large variety of space groups and on structures of different complexity. It may also be applied to small-molecule data without any modification, but we have not made extensive tests in this sense. It is quite likely that the program can be further simplified to gain CPU time without losing efficiency. The most important goal to attain in the near future is to allow the program to be less demanding in the data resolution: we hope that 1.5 Å resolution may be attained, provided some prior information on the general features of the macromolecules can be exploited.

Thanks are due to A. Bacchi, M. Crisma, Z. Dauter, J. C. Fontecilla-Camps, R. C. Haltiwanger, H. Hope, D. A. Langs, P. J. Loll, M. Saviano and C. Weeks who kindly provided us with diffraction data for large structures. A. Soyer is kindly acknowledged for enlightening discussions.

References

Altomare, A., Burla, M. C., Camalli, M., Cascarano, G., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G., Polidori, G. & Spagna, R. (1999). *J. Appl. Cryst.* **32**, 115–119.

Altomare, A., Burla, M. C., Cascarano, G., Giacovazzo, C., Guagliardi, A., Moliterni, A. G. G. & Polidori, G. (1995). *Acta Cryst.* A**51**, 305–309.

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

Bacchi, A., Redenti, E., Amari, G., Delcanale, M., Ventuta, P., Sheldrick, G. M. & Pelizzi, G. (2000). In preparation.

Badger, J. (1997). *Methods Enzymol.* **277**B, 344–352.

Baggio, R., Woolfson, M. M., Declercq, J. P. & Germain, G. (1978). *Acta Cryst.* A**34**, 883–892.

Bhuiya, A. K. & Stanley, E. (1963). *Acta Cryst.* **16**, 981–984.

Burla, M. C., Camalli, M., Carrozzini, B., Cascarano, G., Giacovazzo, C., Polidori, G. & Spagna, R. (1999). *Acta Cryst.* A**55**, 991–999.

Burla, M. C., Cascarano, G., Giacovazzo, C., Lamba, D., Polidori, G. & Ughetto, G. (1999). *Croat. Chem. Acta*, **72**, 519–529.

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

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.

DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.

Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G., Zagari, A. & Mazzarella, L. (2000). *J. Mol. Biol.* In the press.

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

Frazao, C., Soares, C. M., Carrondo, M. A., Pohl, E., Dauter, Z., Wilson, K. S., Hervas, M., Navarro, J. A., De La Rosa, M. A. & Sheldrick, G. M. (1995). *Acta Cryst.* D**51**, 232–234.

Germain G. & Woolfson, M. M. (1968). *Acta Cryst.* B**24**, 91–96.

Gessler, K. (2000). In preparation.

Giacovazzo, C. (1976*a*). *Acta Cryst.* A**32**, 91–99.

Giacovazzo, C. (1976*b*). *Acta Cryst.* A**32**, 100–104.

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

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

Hauptman, H. A. (1991). In *Crystallographic Computing 5: from Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry. IUCr/Oxford University Press.

Hauptman, H. A., Xu, H., Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* A**55**, 891–900.

Hope, H. (1988). *Acta Cryst.* B**44**, 22–26.

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, I. L. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.

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

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

Loll, P. J., Miller, R., Weeks, C. M. & Axelsen, P. H. (1998). *Chem. Biol.* **5**, 293–298.

Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A. & Cambillau, C. (1997). *J. Mol. Biol.* **268**, 779–799.

Merigeau, K., Arnoux, B., Norris, K., Norris, F. & Ducruix, A. (1998). *Acta Cryst.* D**54**, 306–312.

Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.

Nimz, O. (2000). In preparation.

Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.

Prive, G. G., Anderson, D. H., Wesson, L., Cascio, D. & Eisenberg, D. (2000). In preparation.

Schäfer, M., Sheldrick, G. M., Bahner, I. & Lackner, H. (2000). In preparation.

Schäfer, M., Sheldrick, G. M., Schneider, T. R. & Vørtesy, L. (1998). *Acta Cryst.* D**54**, 175–183.

Sheldrick, G. M. (1998). In *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier. Dordrecht: Kluwer Academic Publishers.

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

Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.

Shiono, M. & Woolfson, M. M. (1992). *Acta Cryst.* A**48**, 451–456.

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.

Tong, L., Ho, D. M., Vogelaar, N. J., Schutt, C. E. & Pascal, R. A. Jr (1997). *J. Am. Chem. Soc.* **119**, 7291–7302.

Tronrud, D. E. (1997). *Methods Enzymol.* **277**B, 306–319.

Vojtechovsky, J., Berendzen, J., Chu, K., Schlichting, I. & Sweet, R. M. (2000). In preparation.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.

Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.