Acta Cryst. (1999). A**55**, 991–999

*SIR*99, a program for the automatic solution of small and large crystal structures

Maria Cristina Burla,[b] Mercedes Camalli,[c] Benedetta Carrozzini,[d] Giovanni Luca Cascarano,[d] Carmelo Giacovazzo,[a]* 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.*
E-mail: c.giacovazzo@area.ba.cnr.it

(*Received 5 February 1999; accepted 24 May 1999*)

Abstract

The moduli and flow diagram of the program *SIR*99 are described. New phasing algorithms are proposed working both in direct and in reciprocal space. Their cooperative work is able to solve the structures of both small and large molecules. In particular, small proteins can be solved *ab initio* without any use of prior information and any user intervention. The efficiency of the various algorithms employed by *SIR*99 has been tested, and the role of the tangent formula clarified. The user is also provided with some practical information concerning the computer power needed.

1. Notation

$F_{\mathbf{h}} = |F_{\mathbf{h}}| \exp(i\varphi_{\mathbf{h}})$: structure factor.
$E_{\mathbf{h}} = R_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}})$: normalized structure factor ($R_{\mathbf{h}}$ is its modulus and $\varphi_{\mathbf{h}}$ its phase).
N: number of non-hydrogen atoms in the unit cell.
$D_1(x) = I_1(x)/I_0(x)$, I_i is the modified Bessel function of order i.
$\Phi_3 = \varphi_{\mathbf{h}_1} + \varphi_{\mathbf{h}_2} + \varphi_{\mathbf{h}_3} \quad (\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0)$.
$\Phi_4 = \varphi_{\mathbf{h}_1} + \varphi_{\mathbf{h}_2} + \varphi_{\mathbf{h}_3} + \varphi_{\mathbf{h}_4} \quad (\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 + \mathbf{h}_4 = 0)$.
$C_j = 2|E_{\mathbf{h}} E_{\mathbf{k}_j} E_{\mathbf{h}-\mathbf{k}_j}|/N^{1/2}$.

2. Introduction

Direct methods have definitively solved in practice the phase problem for small molecules. Computer programs like *MULTAN* (Main *et al.*, 1980), *SHELX* (Sheldrick, 1990), *SIR*97 (Altomare *et al.*, 1999) and *SAYTAN* (Debaerdemaeker *et al.*, 1985) solve crystal structures in a more or less routine way and require minimal intervention by the user. Crystal structures with 200 or more non-hydrogen atoms in the asymmetric unit have been rarely solved by the above programs. In particular, small proteins are out of their range, in spite of the pioneering efforts made by Woolfson & Yao (1990) and by Fan *et al.* (1991). The most complete analysis of the problem from a first principles approach has been provided by Giacovazzo *et al.* (1994). Their results may be summarized as:

The α reliability parameter of the classical tangent formula is normally distributed around

$$\langle \alpha_{\mathbf{h}} \rangle = \sum_{j=1}^{r} C_j D_1(C_j)$$

with variance given by

$$\sigma^2_{\alpha_{\mathbf{h}}} \cong \tfrac{1}{2} \sum_{j=1}^{r} C_j^2 [1 + D_2(C_j) - 2D_1^2(C_j)].$$

If, as for usual size proteins, the C_j are very small, then

$$D_1(C_j) = C_j/2$$

$$\langle \alpha_{\mathbf{h}} \rangle = \sum_{j=1}^{r} C_j^2/2$$

$$\sigma^2_{\alpha_{\mathbf{h}}} \cong \langle \alpha_{\mathbf{h}} \rangle.$$

Accordingly, for values of $\langle \alpha_{\mathbf{h}} \rangle < 1$, the signal ($\langle \alpha_{\mathbf{h}} \rangle$) is smaller than the noise ($\sigma_{\alpha_{\mathbf{h}}}$). Under these conditions, the tangent formula will not work efficiently and the correct solution will not be recognized among the various trials.

The result of the analysis was the following statistical solvability criterion: the tangent formula can be successfully applied to a given set of data if, for a sufficiently high percentage of large normalized structure factors,

$$z = \langle \alpha \rangle / \sigma_\alpha > T,$$

where T is a threshold between 2 and 3.

New direct-methods approaches seem to annul the predictive power of the statistical solvability criterion. *Shake-and-Bake* (Weeks *et al.*, 1993) and more recently *Half-Bake* (Sheldrick & Gould, 1995; Sheldrick, 1998) have immoderately enlarged the size of crystal structures solvable by direct methods. The *Shake-and-Bake* (*SnB*) approach differs from the previous computer programs because each trial solution is repeatedly cycled in both real and reciprocal space, and structure (or phase) refinement is performed in each space. As a consequence, the algorithm is computer intensive, and it

has been made feasible only in recent years thanks to the increased computing power of modern computers. *Half-Bake* (*HB*) stresses refinement in direct space.

Pertinent questions which now arise are:

(*a*) Is the statistical solvability criterion wrong?

(*b*) Is greater computer power the main reason for the success of *SnB* and *HB*?

(*c*) Have some new principles been introduced that make solvable a crystal structure even when it is not solvable solely by application of the tangent formula?

(*d*) Are these new principles, if present, well understood and/or previously described?

These questions are of basic importance because correct answers allow one to understand and, more importantly, to improve the present techniques for direct solution of the phase problem. In this paper, we will analyse these questions from a point of view more traditional than that used by *SnB* and *HB*.

(*a*) We will make no use of features that are specific to the structure under examination (*e.g.* presence of disulfide bridges or solvent regions).

(*b*) We will show how and why a traditional program like *SIR*97 can evolve to *SIR*99, a package able to solve macromolecular structures without repeatedly switching from direct to reciprocal space and *vice versa*.

(*c*) An automatic structure refinement will be devised that is able to discard false peaks, recover a satisfactory structural model with reasonable bond lengths and angles and end with sufficiently low *R* values. Of course, such an automatic refinement will be less robust for macromolecular structures, whose final refinement requires significant user intervention (with possible use of geometric restraints and/or constraints). However, our procedure is able to provide a good basis for the final refinement.

3. Structure and flow diagram of *SIR*99

DATA is the data input routine of *SIR*99. *SOLVE* is the modulus devoted to normalizing structure factors, setting up invariant relationships, applying the tangent formula, calculating electron-density maps and performing the preliminary (and automatic) least-squares refinement of the model structure. Access to the graphic interface is available through the *MENU* modulus, and final refinement may be performed through the *LSQ*, *HYDROGEN* and *GEOMETRY* moduli. The last four moduli will not be described here; they do not present any remarkable difference with respect to the corresponding ones in *SIR*97 (Altomare *et al.*, 1999).

After the normalization process, the following steps are executed.

Step 1. Triplet and quartet invariant values are calculated. For the former, the P_{10} formula is used (Cascarano *et al.*, 1984) and quartet invariants are estimated *via* the formula provided by Giacovazzo (1976) and integrated with the supplemental contribution arising from special quintets (Altomare *et al.*, 1995).

Structure seminvariants, psi-zero triplets (Giacovazzo, 1993; Cascarano & Giacovazzo, 1995) and psi-E triplets (Altomare *et al.*, 1991), which were actively used in *SIR*97, have no role in *SIR*99. Suitable tests, not shown for brevity, suggest that, even if useful for some crystal structures, such relationships make the phasing process slower and are of negligible usefulness for large structures.

Step 2. $N_{\rm large}$ reflections (those with the largest $|E|$'s) are selected and phased *via* a double tangent process. Magic integer starting phase sets (see Main, 1978), the default choice of *SIR*97, are not used in *SIR*99. Indeed, random starting phases (Baggio *et al.*, 1978) proved to be of equivalent efficiency for small structures and superior for large ones.

Step 3. The phase values obtained in step 2 are processed by the following three procedures:

(*a*) *EDM* (electron-density modification);

(*b*) *HAFR* (heavy-atom reduced real-space Fourier refinement);

(*c*) *DLSQ* (least squares in diagonal matrix).

It will be clear from later discussion that the real-space refinement techniques used in *SIR*99 introduce the *atomicity* condition in a progressive way. First, Fourier inversion is executed by using single pixels of the electron-density map in *EDM*. The heaviest atomic species (with a suitable occupancy factor and fixed vibrational factor) are represented in *HAFR*. Finally, all the atomic species and their usual refinable parameters are introduced in *DLSQ*.

Step 4. If the crystallographic residual

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

is larger than 0.25, a new random starting set is produced and processed by steps 1–4. Otherwise, the automatic procedure stops and displays a graphical interface allowing interaction with the user.

Step 4 clearly indicates that the figures of merit (FOMs), which played a central role in *SIR*97 to select the correct solution, are no longer used in *SIR*99. Each trial starting set is processed up to the last step, and only the final RES value will indicate if the crystal structure is solved or not. This strategy may be time consuming if the frequency of correct solutions is not very low, but it is able to distinguish solutions even when, as in the case for large structures, the FOMs do not work well.

In spite of the numerous differences outlined above, *SIR*97 and *SIR*99 adopt a similar strategy. The tangent refinement section (reciprocal-space refinement) is followed by the real-space refinement. In *SIR*99, however, the real-space section is much more complex than in *SIR*97 and it provides the additional power needed to solve macromolecular crystal structures.

Table 1. *Code name, space group and crystallochemical data for test structures with* $N_{asym} \leq 100$

Z is the number of molecular formula units in the cell; N_{asym} is the number of non-hydrogen atoms in the asymmetric unit.

Structure code	Space group	Molecular formula	Z	NASYM	Reference
Apapa	$P4_12_12$	$C_{30}H_{37}N_{15}\ O_{16}P_2 \cdot 6H_2O$	8	69	(*a*)
Bcdimp	$P2_1$	$C_{55}H_{76}N_4O_{37}$	2	96	(*b*)
Ergo	$P2_12_12_1$	$C_{28}H_{44}O$	8	58	(*c*)
Goldman2	Cc	$C_{28}H_{16}$	8	56	(*d*)
Jamilas	$P1$	$K_4C_{64}H_{68}N_8O_{20}S_4$	1	100	(*e*)
Mbh2	$P1$	$C_{15}H_{24}O_3$	3	54	(*f*)
Mghex	$P3_1$	$C_{48}H_{68}N_{12}O_{12}Mg \cdot 2ClO_4 \cdot 4CH_3CN$	3	95	(*g*)
Newqb	$P\bar{1}$	$C_{24}H_{20}\ N_2O_5$	4	62	(*h*)
Rc62	$P2_1$	$C_{68}O_{12}$	2	80	(*i*)
Rifolo	$P2_1$	$C_{39}H_{49}NO_{13}CH_3OH \cdot H_2O$	2	53	(*j*)
S6	$P\bar{1}$	$C_{66}H_{70.5}O_{12}N_5S_6$	2	89	(*k*)
Schwarz	$P1$	$C_{46}H_{70}O_{27}$	1	73	(*l*)
Winter2	$P2_1$	$C_{52}H_{83}N_{11}O_{16} \cdot 3CH_2Cl_2$	2	88	(*m*)

References: (*a*) Suck *et al.* (1976); (*b*) Saviano, Iacovino *et al.* (1999); (*c*) Hull *et al.* (1976); (*d*) Irngartinger *et al.* (1981); (*e*) Dodson *et al.* (1990); (*f*) Poyser *et al.* (1986); (*g*) Karle & Karle (1981); (*h*) Grigg *et al.* (1978); (*i*) York University Group, private communication; (*j*) Cerrini *et al.* (1988); (*k*) D. Watkin, private communication; (*l*) B. Schweizer, unpublished; (*m*) Butters *et al.* (1981).

Table 2. *Code name, space group and crystallochemical data for test structures with* $N_{asym} > 100$

Z is the number of molecular formula units in the cell; N_{asym} is the number of non-hydrogen atoms in the asymmetric unit (water molecules excluded).

Structure code	Space group	Molecular formula	Z	N_{asym}	Reference
Alessia	$P2_1\ 2_1\ 2_1$	$C_{119}H_{430}O_{92}N$	4	212	(*n*)
App	$C2$	$C_{190}O_{58}N_{53}Zn$	4	302	(*o*)
C8	$P2_1$	$C_{98}H_{136}O_{16}N_{16}$	2	130	(*p*)
Crambin	$P2_1$	$C_{203}H_{338}O_{65}N_{55}S_6$	2	329	(*q*)
Gramicidin	$P2_1\ 2_1\ 2_1$	$C_{228}H_{370}O_{49}N_{40}$	4	317	(*r*)
Mor59	$P2_1$	$C_{96}H_{164}O_{18}N_{12}$	2	126	(*s*)
Profl	$P\bar{1}$	$C_{128}H_{140}O_{48}N_{40}P_4$	2	220	(*t*)
Rubredoxin	$P2_1$	$C_{243}H_{584}O_{187}N_{57}S_6Fe$	2	494	(*u*)
Toxin II	$P2_1\ 2_1\ 2_1$	$C_{310}H_{703}O_{191}N_{85}S_8$	4	594	(*v*)
Tval	$P1$	$C_{54}H_{90}N_6O_{18}$	1	156	(*w*)
Vancomycin	$P4_3\ 2_1\ 2_1$	$C_{67}H_{79}O_{48}N_9Cl_4$	16	255	(*x*)
X116*a*	$P2_1$	$C_{26}H_{28}NOBr$	2	290	(*y*)
X124	$P1$	$C_{92}H_{76}N_4O_{12}Cl_8S_4$	1	120	(*y*)

References: (*n*) Bacchi *et al.* (1999); (*o*) Glover *et al.* (1983); (*p*) Saviano, Isernia *et al.* (1999); (*q*) Hope (1988); (*r*) Langs (1988); (*s*) Polese *et al.* (1996); (*t*) Burla *et al.* (1999); (*u*) Sheldrick *et al.* (1993); (*v*) Smith *et al.* (1997); (*w*) Karle (1975); (*x*) Loll *et al.* (1997); (*y*) R. C. Haltiwanger, SmithKline Beecham Pharmaceuticals, unpublished.

The above description suggests that the success of *SIR*99 in solving large crystal structures can be better understood if the role of the following tools are analysed:

(*a*) the triplet and quartet formulas;
(*b*) the double tangent procedure;
(*c*) the *EDM*, *HAFR* and *DLSQ* processes.

We have applied *SIR*99 to two sets of crystal structures. The first set (Table 1) contains some structures with fewer atoms than 100 in the asymmetric unit; the second set (Table 2) contains small proteins and non-proteins with more than 100 unique atoms.

4. About the tools of *SIR*99

It will be clear later on that an essential prerequisite for the success of the phasing process is to recognize, when *HAFR* starts, a relatively small subset of reflections with a small phase error from which the phasing pathway can develop. For small structures, the tangent formula is a cheap tool for obtaining such a result even starting from a random set of phases. For macromolecular structures, this is seldom attained. *EDM* and *HAFR* may be considered as supplementary real-space tools devoted to identifying such a suitable set. Having the above observations in mind, we can now examine the efficiency of the various tools of *SIR*99.

4.1. *The* P_{10} *formula*

The main force driving a random phase set to correct values is the reliability of the triplet relationship estimates. Good relationships frequently lead the phasing process to a correct solution; even slight improvements

Table 3. *Rubredoxin, X116a, toxin II: for each structure, the number of triplets (Nr) with Cochran's or P_{10} reliability parameter larger than a fixed ARG is given*

% is the percentage (×100) of triplets with positive cosine and $\langle|\Phi_3|\rangle$ is the average deviation (from zero) of the triplet phases.

		P_3			P_{10}		
Structure code	ARG	Nr	%	$\langle\|\Phi_3\|\rangle$	Nr	%	$\langle\|\Phi_3\|\rangle$
Rubredoxin	0.0	50000	74.8	60.98	50000	74.8	60.98
	0.4	50000	74.8	60.98	45024	75.8	59.72
	0.6	18610	77.1	57.99	19388	78.8	55.77
	0.8	3651	80.1	54.19	5416	82.6	51.24
	1.2	149	83.9	48.11	403	87.8	43.05
X116a	0.0	50000	83.5	49.91	45040	86.0	46.78
	0.4	50000	83.5	49.91	40224	87.7	44.61
	0.6	50000	83.5	49.91	37053	88.7	43.27
	0.8	50000	83.5	49.91	33585	89.8	41.93
	1.2	36407	84.6	48.40	25889	91.8	38.96
	1.6	13637	87.5	44.41	18718	93.4	36.46
	2.6	1096	91.4	37.36	6934	96.0	32.01
	4.4	24	100.0	20.75	856	98.2	25.97
Toxin II	0.0	50000	62.6	75.17	50000	62.6	75.17
	0.2	50000	62.6	75.17	49997	62.6	75.17
	0.4	19366	64.7	72.43	18665	65.0	72.40
	0.6	1513	69.1	66.92	1697	69.3	66.48
	0.8	111	71.2	61.72	131	71.8	60.92

in the estimates can remarkably reduce the computing time necessary for success. The P_{10} formula (Cascarano *et al.*, 1984) estimates triplet phases *via* a von Mises expression, with reliability coefficient

$$G = C(1 + Q), \quad \text{with} \quad Q = \sum_{\mathbf{k}} (A_{\mathbf{k}}/N)/(1 + B_{\mathbf{k}}/N).$$

Each pair $A_{\mathbf{k}}$, $B_{\mathbf{k}}$ arises from the statistical analysis of the special quintet

$$\Phi_5 = \varphi_{\mathbf{h}_1} + \varphi_{\mathbf{h}_2} + \varphi_{\mathbf{h}_3} + \varphi_{\mathbf{k}} - \varphi_{\mathbf{k}}$$

involving the use of ten moduli: their algebraic expressions are not specified here for brevity.

If quintets are not used, G is equivalent to C and P_{10} reduces to the Cochran (1955) formula, referred to here as P_3. In *SIR*99, $\mathbf{k}$ varies at most over the 70 reflections with the largest R values, even for large structures.

The P_{10} formula was checked in *SIR*97 over a large set of small structures. It proved much more efficient than the Cochran formula and was therefore chosen for use in the crystal structure solution of small molecules. P_{10} has never been previously applied to large crystal structures. This is probably because of the following argument. Since the contribution from each quintet Φ_5 to G is of order $N \times N^{1/2}$, one would expect that P_{10} estimates of Φ_3 would be substantially equivalent to Cochran estimates (*i.e.* $G \cong C$). This is not always true. In Table 3, we show, for rubredoxin, X116a and toxin II (three large structures), the mean absolute triplet phase error when calculated according to P_3 and P_{10}. It is clear from this table that the triplet estimates are more efficiently ranked by the P_{10} formula for the first two structures, while no differences are found for toxin II (and, very likely, equal-atom complex structures). The gain is particularly strong for X116a, owing to the presence of the heavy Br atoms.

Table 4. *Rubredoxin, X116a, toxin II: percentage of the negative quartets (%) and average deviation from zero ($\langle|\Phi_4|\rangle$) for quartet phases with reliability factor (as calculated by the SIR97 procedure) larger than ARG*

Structure code	ARG	Nr	%	$\langle\|\Phi_4\|\rangle$
Rubredoxin	0.0	427	53.2	94.27
	0.2	1	100.0	120.00
X116a	0.0	2926	62.0	103.55
	0.2	2124	65.2	107.00
	0.4	1183	67.7	109.93
	0.6	539	69.9	113.37
	0.8	216	68.5	113.04
	1.2	40	82.5	127.35
Toxin II	0.0	24	50.0	75.92
	0.2	0	0.0	0.00

Table 3 suggests that the crystal structure solution of X116a and rubredoxin would be easy, and that toxin II should be resistant. This agrees with the experimental results shown later on in Table 8.

4.2. *The quartet invariants*

Negative estimated quartets are widely (and with different procedures) used in most of the recent direct-methods programs (*e.g. SAYTAN*, *SHELX*, *SnB*, *SIR*97). More recently, *SnB* and *HB* (see Weeks *et al.*, 1999) discouraged their use because they require more CPU time while adding little to the solution of the large structures (the phase information for quartets is of order N^{-1}). This does not fully agree with our experience with *SIR*99: negative quartets still have an important role up to medium-size crystal structures. An overview of the

Table 5. *Minimum mean phase error, for some large test structures, obtained by single-tangent (Min-err1) and double-tangent (Min-err2) procedures, after a fixed number of trials (N_{trials})*

The same number of reflections has been used for the two tangent procedures.

Structure code	N_{trials}	Min-err1	Min-err2
Alessia	1100	80.60	71.65
App	100	46.92	45.49
C8	100	69.98	64.39
Crambin	300	69.71	66.26
Gramicidin	300	78.28	76.29
Mor59	300	66.68	61.03
Rubredoxin	100	43.62	42.67
Toxin II	1100	80.48	67.77
Vancomycin	1100	82.00	78.94

quality of the quartet estimates for rubredoxin, X116a and toxin II is given in Table 4. In agreement with Table 3, the estimates are particularly good for X116a and rubredoxin, and not useful for toxin II. Their numbers are not high and their reliability is small. However, they can play a non-negligible role in increasing the efficiency of the phasing procedure (see Table 8).

4.3. *The double-tangent procedure*

Random starting phases are assigned to $N_{large}/4$ reflections selected by a convergence procedure. The first tangent process extends and refines phases up to $N_{large}/2$ reflections. The values of 40% of the thus-assigned phases (those with the larger α values) are kept fixed in the first cycles of the second tangent process and then refined in the last cycles. The second tangent procedure involves all the N_{large} reflections.

It may be noticed that N_{large} is usually much smaller than the number of reflections phased by the *SnB* and *HB* tangent procedures. Its value is chosen by default in *SIR*99; its maximum value is 1800, attained only for profl, rubredoxin and toxin II. Consequently, the first tangent process is applied to a remarkably small number of reflections: *e.g.* $N_{large}/2 = 368, 359, 205$ for C8, mor59 and ergo, respectively. This choice is suggested by the fact that involving a small number of reflections in a tangent procedure lowers the number of triplet invariants to exploit, but selects (in the average sense) the most reliable ones. Experimental tests (not shown for brevity) confirm that the chance of obtaining reduced phase errors by our tangent program increases when a reduced number of reflections are involved in the first tangent process.

Let us now compare the results obtained by a single-tangent procedure and those obtained *via* a double tangent. Both the procedures involve the same number of reflections (say N_{large}) and the same number of trials (say N_{trials}). In Table 5, we show, for some large structures, the minimum average phase error obtained, for fixed N_{large} and N_{trials} values, by a single-tangent procedure (Min-err1) and by the double-tangent process (Min-err2). It is evident that the double tangent is superior to the single one, so we used it as a default choice for *SIR*99. We were unable to obtain better results by using a triple-tangent procedure.

4.4. *The EDM procedure*

The procedure *FLEX*, recently proposed by Giacovazzo & Siliqi (1997) to introduce solvent-flattening information into the phasing process of proteins (see also Shiono & Woolfson, 1992; Refaat & Woolfson, 1993), inspired the *SIR*99 *EDM* procedure. This procedure consists of 15 supercycles, each constituted of 7 microcycles. In each microcycle, the calculations $\rho \rightarrow \{\varphi\} \rightarrow \rho$ are performed, where ρ represents the electron-density (E) map. In the step $\rho \rightarrow \{\varphi\}$, only a small fraction of the density map is used. In each supercycle, this fraction (which contains the pixels with the largest values of ρ) varies from 2.0% in the first microcycle to 2.5% in the last. The percentage is reset to 2.0% when the next supercycle starts.

It should be noticed that

(i) no use is made of the molecular envelope;

(ii) the phases obtained by Fourier inversion of the modified ρ map are weighted by Sim-like weights;

(iii) the electron-density maps are calculated by using a number of reflections that cyclically increases with the microcycle order (from N_{large} in the first microcycle to the number of reflections with $|E| > 0.8$).

The efficiency of the *EDM* procedure may be estimated from Fig. 1, where we show, for some test structures, the mean phase error *versus* the supercycle order for a trial ending in a correct solution. *EDM* works fine for alessia,

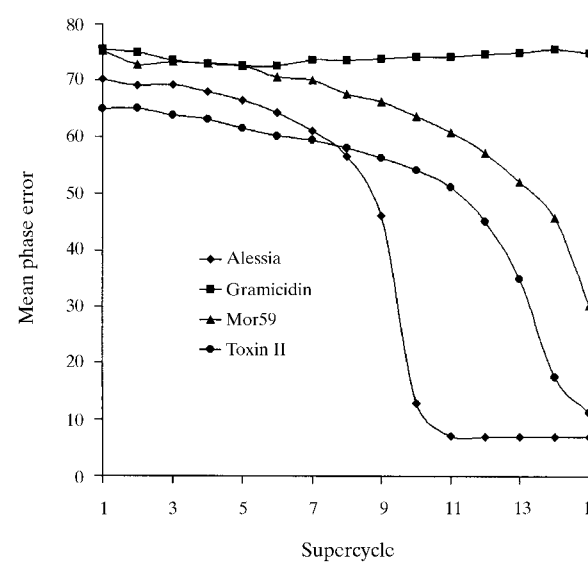

Fig. 1. Mean phase error of some test structures against the number of supercycles in *EDM*.

Table 6. *Values of N_{rinit} and N_{rfin} for all the large test structures*

Structure code	$N_{\rm rinit}$	$N_{\rm rfin}$
Alessia	490	3925
App	670	5365
C8	326	2613
Crambin	725	5802
Gramicidin	700	5605
Mor59	318	2549
Profl	800	6400
Rubredoxin	800	6400
Toxin II	800	6400
Tval	400	3205
Vancomycin	577	4618
X116a	646	5173
X124	329	2629

toxin II and mor59, while gramicidin is insensitive (its phase error will decrease in *HAFR*). It may be worthwhile noting that the sensitivity to *EDM* or *HAFR* procedures seems structure dependent. Fig. 1 suggests that

(i) the double tangent is only able to provide, for large structures, sets of phases with relatively large average errors (from 65 to 75° for the structures in Fig. 1);

(ii) *EDM* extends the phase information and, in favourable cases, reduces the errors;

(iii) the *HAFR* and *DLSQ* steps will contribute to reduce the phase error and to provide the structure model.

The cooperative work of all the steps is, in practice, responsible for the efficiency of the phasing process.

4.5. *The HAFR procedure*

HAFR consists of 37 cycles $\{\varphi\} \rightarrow \rho \rightarrow \{\varphi\}$ (the number was chosen for purely practical reasons). In the first cycle, a quite small number of reflections, say $N_{\rm rinit}$, is automatically selected by the program (those with the largest weight at the end of *EDM*) and used to calculate ρ (denote such a set by $\{\varphi\}_{\rm in}$). The maxima of ρ are arranged in decreasing order of the peak height, and the heaviest atomic species is associated with all the peaks. An occupancy factor is associated with each peak to take into account peak height, site occupancy and chemical connectivity. Then structure factors are calculated using the isotropic thermal factor provided by the Wilson plot, and the cycle $\{\varphi\} \rightarrow \rho \rightarrow \{\varphi\}$ starts again. Two very important features of *HAFR* deserve to be outlined [they make this routine much more efficient than the analogous one used in *SIR*97].

(i) The value of $N_{\rm rinit}$ must be very small even for large structures. Unexpectedly, it does contain sufficient information for solving even large structures. This point is critical when *EDM* ends with a large phase error. Choosing a larger set $\{\varphi\}_{\rm in}$ usually does not result in a solution. We show in Table 6 the value of $N_{\rm rinit}$ for all the large test structures. We also show the values of $N_{\rm rfin}$, the number of reflections to which *HAFR* extends the phasing process in the last cycle.

(ii) The number of peaks selected in the Fourier map cyclically varies between 60 and 80% of $N_{\rm asym}$. Correspondingly, the number of reflections used to calculate the electron density cyclically varies between $N_{\rm rinit}$ and 70% of the measured reflections (those with the largest R value). Without the cyclic variation of the numbers of atoms and reflections, most of the large crystal structures could not be solved by *SIR*99.

The above strategy strongly differs from the algorithms used by *SnB* and *HB* for selecting numbers of peaks and numbers of reflections.

4.6. *The DLSQ procedure*

As in *SIR*97, peaks in the last Fourier map of *HAFR* are labelled in terms of atomic species according to their heights and the chemical content of the cell. Structure-factor calculation, diagonal least-squares refinement and $2F_o - F_c$ Fourier maps are cyclically performed to refine the model structure. Our tests prove that phase errors larger than about 45° do not allow the correct refinement of the structural models.

5. Experimental applications

The P_{10} formula and negative quartet invariants are used despite the opposite advice of *SnB* and *HB* proponents. We will show that the CPU time necessary to calculate thousands of invariants *via* the P_{10} formula is largely compensated by the increased ability of *SIR*99 to find the correct solution. A similar conclusion will also hold for the use of negative quartet invariants. To check the advantage of P_{10} against P_3, we ran *SIR*99 on the test structures quoted in Tables 1 and 2 according to the following protocols.

Protocol 1: The P_{10} formula is used and negative quartet invariants are actively employed. *EDM*, *HAFR* and *DLSQ* are used as described above. This is the default choice of *SIR*99.

Protocol 2: The P_3 formula is used, and negative quartets are not calculated. *EDM*, *HAFR* and *DLSQ* as in protocol 1.

For the set of small structures, *SIR*99 explored a maximum of 300 trials using both protocols, a maximum of 2000 trials were investigated for the set of large structures. The results are shown in Table 7 for the small structures and in Table 8 for the larger ones. TRSOL is the serial number of the trial for which the correct solution was found, RES is the corresponding crystallographic residual for the correct solution and Time is the CPU time (in s for Table 7 and in h for Table 8) sufficient to obtain the quoted value of RES. All tests were performed using a Digital Personal Workstation500au (SPECfp95: 19.5). We observed the following.

Table 7. *Small test structures; for each protocol, TRSOL is the order of the trial for which the correct solution is found, RES is the corresponding crystallographic residual for the correct solution, Time is the CPU time (in s) sufficient to obtain the correct structure model*

	Protocol 1			Protocol 2			Protocol 3		
Structure code	TRSOL	RES	Time	TRSOL	RES	Time	TRSOL	RES	Time
Apapa	37	13.96	3363.1	3	13.91	355.7	6	13.93	807.4
Bcdimp	1	12.77	50.2	14	13.24	852.8	21	13.31	175.5
Ergo	22	15.94	1087.5	69	17.08	3631.1	10	17.23	512.6
Goldman2	1	10.90	38.7	1	10.94	35.3	1	10.91	37.8
Jamilas	1	13.53	60.5	3	13.60	148.9	2	14.04	150.7
Mbh2	1	12.44	17.4	1	12.48	14.1	1	12.41	15.4
Mghex	24	13.00	1633.0	4	12.83	257.6	7	12.84	503.5
Newqb	40	9.50	730.7	82	9.47	1566.8	29	9.48	526.4
Rc62	13	13.77	477.5	6	13.78	209.5	10	12.58	345.8
Rifolo	36	15.48	797.6	9	14.76	204.7	11	14.88	250.7
S6	2	11.82	75.3	300	–	11967.3	206	12.21	8161.8
Schwarz	1	16.61	27.5	1	15.92	23.3	1	16.76	27.2
Winter2	3	23.83	137.2	4	20.6	195.7	15	20.15	712.1
Total time			8496.2			19462.8			12226.9

(*a*) TRSOL changes remarkably with the protocol. If a small number of triplet or quartet relationships are evaluated differently, the phasing pathway is altered completely.

(*b*) The total time necessary to solve the test structures according to protocol 1 is much smaller than that of protocol 2 (see total time at the end of Tables 7 and 8).

(*c*) Some small crystal structures are still unsolved after 300 trials when protocol 2 is applied (too high values of RES for S6). One large structure (alessia) and two proteins (toxin II and vancomycin) are still unsolved after 2000 trials. Thus, the total times required for protocol 2 to solve all the structures in Tables 1 and 2 are actually higher than the values shown.

To check the usefulness of the negative quartet invariants, we ran *SIR*99 according to the following protocol.

Protocol 3: The P_{10} formula is used and negative quartet invariants are not calculated. *EDM*, *HAFR* and *DLSQ* as in protocol 1. This protocol estimates the relative usefulness of the quartet invariants.

The results in Tables 7 and 8 indicate that

(*a*) the additional use of the negative quartet relationships can deeply change the phasing pathway relative to protocol 1 (RES is often different for protocols 1 and 3);

(*b*) for the small structures, the total time related to protocol 3 is larger than for protocol 1, but smaller than for protocol 2 (and now S6 is solved in the first 300 trials);

(*c*) for large structures, the total times for protocols 1 and 3 are very close because quartets have little effect on the phasing process for large structures.

The above results clearly indicate that the use of P_{10} and of the negative quartet invariants is a necessary ingredient for the success of the *SIR*99 recipe. Furthermore, any new probabilistic estimate of the triplet and negative quartet invariants should be considered an important tool for improving the efficiency of direct procedures.

6. *SIR*99 *versus SIR*97

The strategy of *SIR*97 may be summarized as follows.

(*a*) A tangent formula including P_{10} triplet estimates, negative quartet invariants and psi-zero relationships is applied to an automatically predetermined number of starting set of phases obtained by magic integer techniques.

(*b*) The correct solution is sought among the most probable ones selected by suitable figures of merit.

(*c*) If no solution is found, the tangent formula is applied to larger starting sets (magic integer approach).

(*d*) If still no solution is found, random starting sets are used (300 by default) and the correct solution sought among the most probable ones.

The description of the complex real-space algorithms of *SIR*99 made above suggests that, in contrast to *SIR*97, *SIR*99 is a computer-intensive program. Its major advantage lies in its ability to solve macromolecular structures completely inaccessible to *SIR*97. The question then arises: Is *SIR*99 wasting CPU time for small molecules? To obtain a realistic knowledge of the relative performances, we show in Table 9, for *SIR*97, the RES value corresponding to the correct solution and the CPU time (Time) necessary to find it.

If we compare the total time in Table 9 with the total time in Table 7 (protocol 1), we see that *SIR*97 requires half of the time to solve the set of small crystal structures. The obvious conclusion is that the present version of *SIR*99 is not competitive with *SIR*97 for the solution of small molecules. It is very likely that the *SIR*99 procedure can be simplified for small structures (in some ways, we are now shooting fish with a cannon). We

Table 8. *Large test structures: for each protocol, TRSOL is the order of the trial for which the correct solution is found, RES is the corresponding crystallographic residual for the correct solution, Time is the CPU time (in h) sufficient to obtain the correct structure model*

	Protocol 1			Protocol 2			Protocol 3		
Structure code	TRSOL	RES	Time	TRSOL	RES	Time	TRSOL	RES	Time
Alessia	797	14.51	95.05	2000	–	205.97	2000	–	233.07
App	19	19.27	4.04	111	19.60	25.52	9	20.01	1.93
C8	15	15.21	0.44	39	15.14	1.26	10	15.22	0.31
Crambin	190	15.10	40.69	86	15.09	23.22	93	15.17	19.23
Gramicidin	68	20.59	14.76	179	20.33	39.33	228	20.40	48.95
Mor59	195	16.32	5.21	42	16.16	1.23	340	16.00	9.23
Profl	2	17.48	0.27	1	17.31	0.19	2	17.48	0.27
Rubredoxin	15	17.23	4.42	5	17.32	1.40	15	17.23	3.94
Toxin II	1024	19.80	567.30	2000	–	1065.62	1024	19.80	567.30
Tval	1	11.44	0.02	1	11.43	0.02	1	11.44	0.02
Vancomicin	909	18.08	414.57	2000	–	839.23	909	18.08	414.57
X116a	1	11.54	0.11	30	11.57	3.27	2	11.49	0.19
X124	1	10.74	0.02	1	10.65	0.02	1	10.74	0.02
Total time			1126.90			2206.29			1279.03

Table 9. *SIR97 performances on the set of small test structures*

RES is the crystallographic residual for the correct solution, Time is the CPU time (in s) sufficient to obtain the correct structure model.

Structure code	RES	Time
Apapa	14.85	49.5
Bcdimp	14.68	2047.0
Ergo	17.55	385.7
Goldman2	10.97	17.2
Jamilas	14.72	215.2
Mbh2	13.72	12.3
Mghex	13.21	654.2
Newqb	10.53	14.3
Rc62	13.50	306.6
Rifolo	17.97	17.3
S6	15.30	36.2
Schwarz	16.14	25.7
Winter2	22.66	920.3
Total Time		4701.5

Table 10. *For some large test structures, the minimum mean phase error (over N_{large} reflections), as obtained by a single tangent after 3000 trials [Min-err1(3000)] and after 10000 trials [Min-err1 (10000)]*

Structure code	Min-err1 (3000)	Min-err1 (10000)
Alessia	78.83	75.06
App	45.92	45.44
C8	64.69	61.02
Crambin	66.35	64.66
Gramicidin	71.99	71.89
Mor59	56.65	32.22
Rubredoxin	42.87	40.99
Toxin II	76.66	74.18
Vancomycin	81.33	80.50

intend to introduce such modifications in the near future.

7. About the role of the tangent procedures

The efficiency of the tangent formula in solving the phase problem for small structures is not questioned. On the contrary, its role for solving large molecules is still dubious. We showed in Table 5 and in Fig. 1 that the application of single- or double-tangent procedures to sets of random phases ended with phase errors still quite large even for those trials for which the correct solution was later attained. No structural information could be directly obtained from the phases assigned by tangent refinement; furthermore, no figure of merit (among those we checked) was able to recognize the trials leading to the correct solution.

One might wonder whether a small mean phase error could be obtained by the tangent process by just enlarging the number of explored trials. In this case, a different strategy could be applied: increase the number of trials and omit (or reduce the complexity of) *EDM* and *HAFR*. We applied a single-tangent procedure to 3000 and 10 000 trials for most of the large structures and the minimum mean phase errors for each test structure are shown in Table 10. If we compare such errors with those quoted in Table 5, we conclude that more favourable phase sets can be produced by the tangent procedure by enlarging the number of trials but the minimum mean phase error does not remarkably decrease for the large structures. This clearly indicates that most of the phasing power of *SIR*99 is in *EDM*, *HAFR* and *DLSQ* steps, rather than in the tangent process.

A second question arises: Can we succeed if we omit the tangent process from *SIR*99 and directly apply *EDM* to random phases? It may be noticed that several phasing algorithms do not use tangent procedures: *e.g.* the Metropolis technique suggested by Bhat (1990), the simulated-annealing procedure proposed by Su (1995), the *RAGA* process by Ramachandran (1990), the forced coalescence method proposed by Drendel *et al.* (1995).

A tangent-less version of *SIR*99 is in preparation. Its optimization requires several additional changes in *EDM* and *HAFR*. We anticipate that for small structures such a version would work with an efficiency comparable with *SIR*97 but it would lose efficiency for large molecules. In conclusion, it seems that we can celebrate the funeral ceremonies of the tangent formula for small crystal structures soon after we have recognized that it had solved the phase problem, but still we have to consider its performances for large molecules for which it was considered less useful.

We thank the IRMEC Institute which, sharing an interest in the solution of small crystal structures, allowed one of us (GLC), together with BC, to take care of aspects concerning these structures.

Thanks are also 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 the diffraction data for the large structures.

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.

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

Bacchi, A., Redenti, E., Amari, G., Delcanale, M.,Ventuta, P., Sheldrick, G. M. & Pelizzi, G. (1999). In preparation.

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

Bhat, T. N. (1990). *Acta Cryst.* A**46**, 735–742.

Burla, M. C., Cascarano, G., Giacovazzo, C., Lamba, D., Polidori, G. & Ughetto, G. (1999). *Croatica Chem. Acta.* Submitted.

Butters, T., Hütter, P., Jung, G., Pauls, N., Schmitt, H., Sheldrick, G. M. & Winter, W. (1981). *Angew. Chem.* **93**, 904–905.

Cascarano, G. & Giacovazzo, C. (1995). *Acta Cryst.* A**51**, 820–825.

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

Cerrini, S., Lamba, D., Burla, M. C., Polidori, P. & Nunzi, A. (1988). *Acta Cryst.* C**44**, 489–495.

Cochran, W. (1955). *Acta Cryst.* **8**, 473.

Debaerdemaeker, T., Tate, C. & Woolfson, M. M. (1985). *Acta Cryst.* A**41**, 286–290.

Dodson, C., Fattah, J., Prout, C. K., Téyman, J. M. & Watkin, D. J. (1990). Personal communication.

Drendel, W. B., Dave, R. D. & Jain, S. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 547–554.

Fan, H. F., Hao, Q. & Woolfson, M. M. (1991). *Z. Kristallogr.* **197**, 197–208.

Giacovazzo, C. (1976). *Acta Cryst.* A**32**, 91–99, 100–104.

Giacovazzo, C. (1993). *Z. Kristallogr.* **206**, 161–171.

Giacovazzo, C., Guagliardi, A., Ravelli, R. & Siliqi, D. (1994). *Z. Kristallogr.* **209**, 136–142.

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.

Grigg, R., Kemp, J., Sheldrick, G. M. & Trotter, J. (1978). *J. Chem. Soc. Chem. Commun.* pp. 109–111.

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

Hull, S. E., Leban, I., Main, P., White, P. S. & Woolfson, M. M. (1976). *Acta Cryst.* B**32**, 2374–2381.

Irngartinger, H., Reibel, W. R. K. & Sheldrick, G. M. (1981). *Acta Cryst.* B**37**, 1768–1771.

Karle, I. L. (1975). *J. Am. Chem. Soc.* **97**, 4379–4386.

Karle, I. L. & Karle, J. (1981). *Proc. Natl Acad. Sci. USA*, **78**, 681–685.

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.

Main, P. (1978). *Acta Cryst.* A**34**, 31–38.

Main, P., Fiske, S. J., Hull, S. E., Lessinger, L., Germain, G., Declercq, J. P. & Woolfson, M. M. (1980). *MULTAN*80*: a System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data.* Universities of York, England, and Louvain, Belgium.

Polese, A., Formaggio, F., Crisma, M., Valle, G., Toniolo, C., Bonora, G. M., Broxterman, Q. B. & Kamphuis, J. (1996). *Chem. Eur. J.* **2**, 1104–1111.

Poyser, J. R., Edwards, R. L., Anderson, J. R., Hursthouse, M. B., Walker, N. C. P., Sheldrick, M. G. & Whalley, A. J. S. (1986). *J. Antibiot.* **39**, 167.

Ramachandran, G. N. (1990). *Acta Cryst.* A**46**, 359–365.

Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.

Saviano, M., Iacovino, R., Benedetti, E., Pedone, C., Impellizzieri, G., Pappalardo, G. & Rizzarelli, E. (1999). In preparation.

Saviano, M., Isernia, C., Rossi, F., Di Blasio, B., Iacovino, R., Mazzeo, M., Pedone, C. & Benedetti, E. (1999). *Biopolymers.* Submitted.

Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.

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.

Su, W. P. (1995). *Acta Cryst.* A**51**, 845–849.

Suck, D., Manor, P. C. & Saenger, W. (1976). *Acta Cryst.* B**32**, 1727–1737.

Weeks, C. M., De Titta, G. T., Miller, R. & Hauptman, H. A. (1993). *Acta Cryst.* D**49**, 179–181.

Weeks, C. M., Sheldrick, G. M., Miller, R., Usón, I. & Hauptman, H. A. (1999). *Bull. Czech Slovak Crystallogr. Assoc.* Submitted.

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