Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Ab initio phasing at resolution higher than experimental resolution

Rocco Caliandro,[a] Benedetta Carrozzini,[a] Giovani L. Cascarano,[a] Liberato De Caro,[a] Carmelo Giacovazzo[a,b]* and Dritan Siliqi[a]

[a]Istituto di Cristallografia, CNR, Via G. Amendola 122/o, 70126 Bari, Italy, and [b]Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125, Bari, Italy

Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

Received 15 February 2005
Accepted 16 May 2005

Owing to the limited experimental resolution of data in macromolecular crystallography, *ab initio* phasing is successful only when atomic or quasi-atomic resolution data are available. It is shown that extrapolating the moduli and phases of non-measured reflections beyond and behind the experimental resolution limit makes the *ab initio* phasing process more efficient and leads to crystal structure solution even in cases in which the standard *SIR*2004 program does not succeed. Moreover, use of the extrapolated values improves the quality of the final electron-density maps and makes the recognition of the correct structrure among several trial structures easier.

1. Notation

NMRE: non-measured reflection extrapolation.
EDM: electron-density modification.
$\mathrm{CORR_{exp}}$: correlation of the trial electron-density map with the true (published) one within the resolution limit of the experimental data. The NMRE procedure is not used.
$\mathrm{CORR_{extra}}$: correlation of the trial electron-density map with the true (published) one within the resolution limit of the experimental data. The NMRE procedure is used.
DM: direct methods.
PM: Patterson methods.
DSR: direct-space refinement performed *via* electron density-modification techniques. In *SIR*2004, DSR follows the application of DM or PM.
$\mathrm{RES_{obs}}$: resolution limit of the experimental data in Å.
$\mathrm{RES_{ext}}$: resolution limit of the reflections estimated by the NMRE procedure.
MPE: mean phase error of a given set of phases with respect to the refined (published) model.
MIS: percentage of non-measured reflections up to $\mathrm{RES_{obs}}$.
Residues: number of residues in the protein molecule.
fFOM: final FOM (fFOM > 1 for the correct solution).
f: atomic scattering factor, thermal factor included.
$|E_{\mathbf{h}}^{\mathrm{obs}}|$: normalized value of the observed structure-factor modulus.
$|E_{\mathbf{h}}^{\mathrm{ext}}|$: normalized value of the structure-factor modulus assigned by the NMRE procedure to extrapolated reflections.
$|E_{\mathbf{h}}^{\mathrm{calc}}|$: normalized structure-factor modulus calculated by inversion of the current electron-density map.
$D_1(x) = I_1(x)/I_0(x)$, where $I_i(x)$ is the modified Bessel function of order i.

2. Introduction

One of the most important limits for *ab initio* phasing in macromolecular crystallography is the limited resolution of

the available experimental data. It is a common belief that atomic (say RES_{obs} = 1 Å) or quasi-atomic (say RES_{obs} = 1.2 Å) resolution is a necessary condition for solving protein crystal structures (Sheldrick, 1990). If this condition is not obeyed, DM would hardly ever provide sets of promising starting phases and DSR would rarely improve them. Accordingly, the atomic resolution condition is assumed to be the working rule in all papers describing the most documented programs for *ab initio* phasing (unless some supplementary information is available): *SnB* (Weeks *et al.*, 1994; Rappleye *et al.*, 2002), *SHELXD* (Sheldrick, 1998), *ACORN* (Foadi *et al.*, 2000) and *SIR*2002 (Burla, Carrozzini, Cascarano, Giacovazzo & Polidori, 2002). However, a recent paper (Burla *et al.*, 2003) suggested that under favourable circumstances it is possible to relax the rule: the new procedures extended RES_{obs} to 1.4–1.5 Å and were implemented in the program *SIR*2004 (Burla *et al.*, 2005).

In spite of this success, the resolution remains the most important factor limiting the solution of the phase problem. In a recent paper (Caliandro, Carrozzini, Cascarano, De Caro, Giacovazzo, Moustiakimov *et al.*, 2005), a new algorithm was suggested to reduce the negative effects of the finite data resolution: the procedure, called NMRE (non-measured reflection extrapolation), involves the use of extrapolated moduli and the phases of non-measured reflections (with resolution lower or higher than the experimental resolution) in classical EDM techniques. The procedure (for early approaches, see Karle & Hauptman, 1964; Seeman *et al.*, 1976; Langs, 1998) was implemented in a modified version of the program *SIR*2004 and applied to electron-density maps obtained under the following conditions: (i) from *ab initio* phasing, RES_{obs} in the interval 1.5–1.0 Å, an approximated electron density available, with MPE in the range (25, 60°) and (ii) from SAD-MAD, SIR-MIR and SIRAS-MIRAS phases, RES_{obs} in the interval 2.8–1.5 Å, an approximated electron density available (*e.g.* after the application of EDM procedures), with MPE in the range (40, 65°).

The applications of the new procedure clearly show that in both cases (i) and (ii) the electron-density maps are more interpretable and resolved than those obtainable *via* the measured reflections only. In the present work, we investigate the usefulness of the NMRE procedure in another typical case of interest in macromolecular crystallography: *ab initio* phasing, with RES_{obs} in the interval 1.5–1.0 Å and no prior phase information available.

Since DM and PM are typically multisolution approaches, numerous trials are usually exploited to solve the phase problem. The use of the experimental data can generate three typical situations: (i) all the trials are characterized by large phase errors (MPE between 80 and 90°) and the corresponding electron-density maps are practically uncorrelated with the true one, (ii) some trials show an MPE in the range (60, 80°) and the corresponding electron-density maps are weakly correlated with the true map and are therefore not interpretable or (iii) one or more trials are characterized by small MPE values, the corresponding electron-density maps are interpretable and the phase problem is solved. We will show that the use of the NMRE algorithm improves the results obtainable *via* the observed reflections only. We have implemented the NMRE approach in a modified version of *SIR*2004 and we have applied it to some test structures.

3. The algorithm

The phasing strategy of *SIR*2004 for macromolecules may be summarized as follows.

(i) Depending on circumstances, PM or DM provides initial trial phases: an early figure of merit (eFOM) selects the most promising trials, which are submitted to DSR (Burla *et al.*, 2003). This combines cycles of EDM with cycles of HAFR (a selected number of large-intensity electron-density peaks are expressed in terms of the heaviest atomic species and of suitable occupancy factors) and of LSQH (the isotropic displacement parameters of the heavy atoms are refined *via* a least-squares procedure).

(ii) The phases provided at the end of the DSR process are often far from the correct ones. Therefore, as a default, we have iterated the DSR procedure (up to 30 iterations for the PM trials and up to five for DM trials) by using as starting phases a few thousand current phase values: those with the largest weights. This iterative process, although time-consuming, succeeds in many difficult cases (*e.g.* protein structures diffracting to 1.4–1.5 Å). In our notation, the first DSR process will be denoted as iteration zero.

(iii) The correct solution is identified by the combined figure of merit fFOM (Burla *et al.*, 2005). Values of fFOM larger than unity suggest that an interpretable electron-density map should be available. The DSR iterations automatically stop when fFOM > 1.0 and the program waits for further user commands; in their absence, another iteration starts. Usually, fFOM < 1 for maps with CORR $\leq$ 0.6. When the map improves by attaining a CORR value larger than 0.7 then fFOM is usually $\geq$1 and it reaches very large values for atomic resolution data when CORR $\geq$ 0.8.

In the modified version of *SIR*2004, the moduli and phases of unobserved reflections are extrapolated after Fourier inversion of the current electron-density map. Preliminary tests have shown that the process is useful only if the corresponding CORR value is sufficiently high. Therefore, we have found that it is convenient to extrapolate unobserved reflections only at iteration one of the DSR process, including non-measured reflections lying below RES_{obs}. If $RES_{obs} \leq 1.2$ Å we extrapolate up to RES_{ext} = 0.8 Å and if RES_{obs} is in the interval 1.3–1.5 Å we extrapolate unobserved reflections up to RES_{ext} = 1.2 Å. Both these resolution thresholds have been found as a result of several tests directed to find optimal RES_{ext} values.

In accordance with Langs (1998) and Caliandro, Carrozzini, Cascarano, De Caro, Giacovazzo, Moustiakimov *et al.* (2005), in the half-cycles $\rho \rightarrow \varphi$ (structure-factor calculation by Fourier inversion of the electron-density map) we found it advantageous to extrapolate all non-measured reflections in one step from RES_{obs} to RES_{ext}, rather than to increase the extrapolation resolution gradually. However, not all the

extrapolated reflections are used in the half cycles $\varphi\rightarrow\rho$ (electron-density map calculation by using suitably selected reflections), but only a small percentage of them, which increases with the DSR macrocycle number (it is equal to 2% of all non-measured reflections in the first DSR macrocycle and to 15% in the subsequent ones). The rationale for such strict limits is the following. The extrapolated reflections actively used in the half cycle $\varphi\rightarrow\rho$ are selected on the basis of their moduli $|E_{\mathbf{h}}^{\rm ext}|$ (as calculated in §4). If the moduli and the phases of such reflections are wrongly estimated, the electron-density map will be less useful than that calculated *via* the observed reflections only and this will damage subsequent extrapolations. Accordingly, an excessive number of extrapolated reflections actively used in the half cycle could easily corrupt the poor information contained in the initial electron-density map and lead to failure of the phasing process.

Other features of the NMRE procedure are the following.

(i) In the half cycles $\rho\rightarrow\varphi$ only a fraction of ρ corresponding to 10% of the volume occupied by the protein is used in each Fourier inversion of the map.

(ii) In the half cycles $\varphi\rightarrow\rho$ the Fourier coefficients are $|E_{\mathbf{h}}^{\rm obs}|$ for observed reflections and $|E_{\mathbf{h}}^{\rm ext}|$ (as calculated in §4) for extrapolated reflections. The associated phase values are those calculated by Fourier inversion of the electron-density map.

(iii) The $|E_{\mathbf{h}}^{\rm calc}|$ values obtained after each map inversion are modified *via* histogram matching to fit the distribution of normalized structure factors expected for a random-atom structure.

(iv) In the half cycles $\varphi\rightarrow\rho$ a Sim-like weight w is associated with each reflection: $w_{\mathbf{h}} = D_1(k|E_{\mathbf{h}}^{\rm obs}||E_{\mathbf{h}}^{\rm calc}|)$ for an observed reflection and $w_{\mathbf{h}} = D_1(k|E_{\mathbf{h}}^{\rm ext}|^2)$ for an extrapolated reflection. k is an empirical constant set to 0.5.

4. About the value of $E_{\mathbf{h}}^{\rm ext}$

Srinivasan & Ramachandran (1965) [see also Pannu & Read (1996) and Caliandro, Carozzini, Cascarano, De Caro, Giacovazzo & Siliqi (2005)] derived, for a structure of N atoms in $P1$, the conditional probability of the observed structure factor modulus R when p atoms have been located with or without errors in the coordinates. Let R_p be the modulus of such a structure factor. The corresponding distribution $P(R|R_p)$ is applied in maximum-likelihood refinement of macromolecular structures (Lunin & Urzhumtsev, 1984; Pannu & Read, 1996; Murshudov *et al.*, 1997; de La Fortelle & Bricogne, 1997),

$$P(R|R_p) = \frac{2R}{(e-\sigma_A^2)}\exp\left[-\frac{1}{(e-\sigma_A^2)}(R^2+\sigma_A^2R_p^2)\right]I_0(X), \quad (1)$$

where

$$e = (1+\sigma_R^2),$$
$$\sigma_R^2 = \langle|\mu|^2\rangle/\Sigma_N,$$

μ represents the experimental error of the structure-factor modulus, ε is the correction factor for expected intensities in reciprocal lattice zones (from Wilson statistics),

$$\Sigma_N = \varepsilon\sum_{j=1}^{N} f_j^2,$$
$$\Sigma_p = \varepsilon\sum_{j=1}^{p} f_j^2$$
$$X = \frac{2\sigma_A R R_p}{(e-\sigma_A^2)}$$
$$\sigma_A \simeq \left(\frac{\Sigma_p}{\Sigma_N}\right)^{1/2} D$$

and

$$D = \langle\cos(2\pi\mathbf{h}\Delta\mathbf{r})\rangle$$

is the mean error on the coordinates of the located atoms. The value of σ_A may be calculated *via* the equation

$$\langle R^2R_p^2\rangle = (e+\sigma_A^2),$$

where the average is calculated by dividing the observed $\sin\theta/\lambda$ interval into resolution shells.

(1) may also be used for extrapolating the moduli of the unobserved structure factors. In this case e has to be set to unity. Furthermore, the σ_A values for extrapolated reflections are obtained by using the least-squares line suggested by the observed $\sin\theta/\lambda$ range. From (1) we obtain (see also Pannu & Read, 1996),

$$\langle R|R_p\rangle = \frac{\pi^{1/2}}{2}(1-\sigma_A^2)^{1/2}\,{}_1F_1\left[-\frac{1}{2};1;-\frac{\sigma_A^2R_p^2}{(1-\sigma_A^2)}\right], \quad (2)$$

where ${}_1F_1$ is the confluent hypergeometric function and

$$\langle R^2|R_p\rangle = 1+\sigma_A^2(R_p^2-1). \quad (3)$$

Introducing the approximation (see Appendix A)

$${}_1F_1(-\frac{1}{2};1;-z^2) \simeq \left(1+\frac{4z^2}{\pi}\right)^{1/2}$$

gives

$$\langle R|R_p\rangle = \frac{1}{2}\left[\pi(1-\sigma_A^2)+4\sigma_A^2R_p^2\right]^{1/2}. \quad (4)$$

Let us now estimate the variance associated with the expectations. We have

$$v_1 = \langle R^2|R_p\rangle - \langle R|R_p\rangle^2 = \frac{3\pi}{4}(1-\sigma_A^2).$$

Since one can also use (3) [instead of (2)] to extrapolate structure-factor moduli, we calculate the relative variance

$$\langle R^4|R_p\rangle = 2(1-\sigma_A^2)^2+4\sigma_A^2R_p^2(1-\sigma_A^2)+\sigma_A^4R_p^4$$

from which

$$v_2 = \langle R^4|R_p\rangle - \langle R^2|R_p\rangle^2 = (1-\sigma_A^2)^2+2\sigma_A^2(1-\sigma_A^2)R_p^2.$$

We note the following.

(i) If $\sigma_A = 0$ there is no correlation between the p-model substructure and the structure. In this case the first does not

Table 1
PDB codes of the 63 test structures: when not available, a code name and reference are specified.

The structures are grouped according to RES_{obs}. For each group we specify (i) the structures solvable by *SIR*2004 and the number of DSR iterations necessary to attain the solution if different from zero and (ii) the structures unsolvable by *SIR*2004 but solvable if the NMRE procedure is used.

$RES_{obs} \leq 1.2$ Å	
Solved by *SIR*2004 without DSR iterations	1a7z; ALESSIA (Bacchi *et al.*, 2002); APP (Glover *et al.*, 1983); 1a6k; 1fy2; 1exr; 2knt; 1a0m; CRAMBIN (Weeks *et al.*, 1995); 1cex; 1ctj; 1c75; 1ick; 2fdn; FIN (courtesy of O. Nimz); 1i76; GRAMICIDIN (Langs, 1988); 1b0y; 1cku; 1bx7; 1dy5; JOD (courtesy of O. Nimz); 1mso; 1b9o; LYSOZYME (Deacon *et al.*, 1998); 1a6m; 1eb6; 1mfm; 2pvb; 2erl; 1kf3; 8rxn; 1irn; 1iro; 1aho; 1sho; 1aa5; 1bx7; 1hhy; TRIVANCO (Loll *et al.*, 1998).
Solved by *SIR*2004 by at least one DSR iteration	1bkr, 4; 3pyp, 1; 1gmx, 2; 1igd, 1; 1a6g, 7
Unsolved by standard *SIR*2004, but solved if the NMRE procedure is implemented	1byz, 4; 1nkd, 2; 1d4t, 4; 1a6n, 8; 352d, 23
$RES_{obs} > 1.2$ Å	
Solved by *SIR*2004 without DSR iterations	9pti (1.22 Å); 1e29 (1.21 Å); 1bx8 (1.38 Å); 1ix2 (1.54 Å); 1awd (1.40 Å)
Solved by *SIR*2004 with at least one DSR iteration	1aac, 2 (1.31 Å); 1lri, 3 (1.45 Å); 3ebx, 3 (1.40 Å); 193l, 3 (1.33 Å); 1paz, 1 (1.55 Å); 1fs3, 1 (1.35 Å); 1ccr, 17 (1.5 Å)
Unsolved by standard *SIR*2004, but solved if the NMRE procedure is implemented	1dxd, 24 (1.40 Å)

provide any useful information on the second: the expected value of R is $\pi^{1/2}/2$ according to (2) and is 1 according to (3), as would be expected from Wilson's distribution. Both ν_1 and ν_2 are independent of the R and R_p values (they are equal to $3\pi/4$ and 1, respectively).

(ii) If $\sigma_A = 1$ (the *p*-substructure coincides with the structure), then $R = R_p$ for both (2) and (3). In this case the variance vanishes, as would be expected.

(iii) If (2) is used, R is always closer to $(\pi/2)^{1/2}$ (the Wilson expectation) than R_p. Analogously, if (3) is used, R is always closer to 1 than R_p. This trend is stronger for decreasing values of σ_A. We show in Fig. 1 the trend of R against R_p [according to (2) and (3)] for the cases in which $\sigma_A = 0.3$ and 0.7. (2) and (3) do not provide equivalent results: one overestimates (while the other underestimates) expectations.

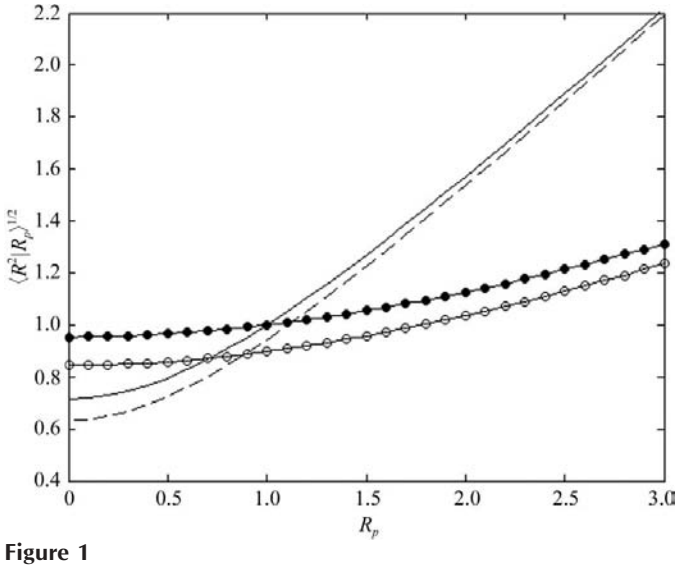

Figure 1
Trend of R against R_p according to (2) when $\sigma_A = 0.3$ (circles) and $\sigma_A = 0.7$ (dotted line) and (3) when $\sigma_A = 0.3$ (filled circles) and $\sigma_A = 0.7$ (unbroken line). See §4 for details.

In our extrapolation procedure we can use as $|E_{\mathbf{h}}^{\rm ext}|$ the square root of the expected R^2 value provided by (3) as well as the value provided by (2). In practice, we should replace (2) by its approximation (4). In our calculations we preferred to use as $|E_h^{\rm ext}|$, the square root of the expected R^2 value provided by (3).

It is worthwhile noting the following. (i) The distribution $P(R|R_p)$ for centric reflections does not coincide with (1), but the value of $\langle R^2|R_p\rangle$ for centric and acentric reflections is the same (Pannu & Read, 1996>). Therefore, no special treatment is necessary for centric reflections. (ii) $|E_h^{\rm ext}|$ is always closer to 1 than $|E_{\mathbf{h}}^{\rm calc}|$.

The set of extrapolated reflections actively used in the half cycles $\varphi \rightarrow \rho$ is selected on the basis of the ratio $|E_{\mathbf{h}}^{\rm ext}|/\nu_1$. Indeed, reflections with maximum ratio $|E_{\mathbf{h}}^{\rm ext}|/\nu_1$ are (on average) those with large $|E_{\mathbf{h}}^{\rm ext}|$ values (*i.e.* the most informative for a Fourier synthesis) and with smaller uncertainty in the estimated moduli.

5. Applications

The efficiency of *SIR*2004 was checked by applying it to 47 macromolecular test structures (Burla *et al.*, 2005): all were solved by using one or more iterations. To check the NMRE procedure, the set of test structures was enlarged to 63, including some data sets that were unsolvable by *SIR*2004: in Table 1 we give their PDB codes (when not available, a code name and reference are specified). The structures are grouped according to RES_{obs}. For each group we specify (i) the structures solved by *SIR*2004 and the number of DSR iterations necessary to obtain the solution if different from zero and (ii) the structures unsolved by the standard *SIR*2004 but solved if the NMRE procedure is used and the number of iterations necessary for their solution.

Table 2
Test structures which are unsolvable or cannot be solved by the standard *SIR*2004 at the iteration zero: they are sorted with respect to RES_{obs}.

In addition to the structure code, the following information is given: PDB is the file code in the Protein Data Bank, Residues is the number of residues per molecule, MIS is the percentage of missing reflections below the experimental resolution limit; the heavy-atom content and the *B* factor obtained by the Wilson plot are also given.

Structure code	PDB	Residues	RES_{obs} (Å)	MIS	Heavy atoms	*B* (Å^2)	Reference
PYP	3pyp	125	0.86	1.6	S_6	6.3	Genick *et al.* (1998)
ALPHA1	1byz	52	0.90	15.0	Cl	3.6	Privé *et al.* (1999)
TETRAPLEX	352d	96	0.95	10.1	Na_{14}, Ca_9	5.1	Phillips *et al.* (1997)
GLPE	1gmx	108	1.06	0.0	S_6, Na	9.6	Spallarossa *et al.* (2001)
METAXIA	1nkd	65	1.10	1.8	S_4	9.8	Vlassi *et al.* (1998)
SH2	1d4t	73	1.10	13.0	S_3	9.5	Lewis *et al.* (1999)
CALPO	1bkr	109	1.10	1.1	S_4	7.2	Banuelos *et al.* (1998)
PROTG	1igd	61	1.10	5.0	S	7.3	Derrick & Wigley (1994)
CARBO	1a6g	151	1.14	9.0	S_4,Fe	9.8	Vojtechovsky *et al.* (1999)
DEOXY	1a6n	151	1.14	2.7	S_4,Fe	8.8	Vojtechovsky *et al.* (1999)
CUPRE	1aac	105	1.30	2.3	S_6,Cu	9.5	Durley *et al.* (1993)
HEWL133	193l	129	1.33	9.0	S_{10}, Cl, Na	14.7	Vaney *et al.* (1996)
WILD	1fs3	124	1.35	1.5	S_{12}	10.9	Chatani *et al.* (2002)
ERBUTOX	3ebx	62	1.40	0.7	S_9	11.6	Smith *et al.* (1988)
MB20	1dxd	154	1.40	2.0	S_4, Fe	9.6	Brunori *et al.* (2000)
COLE	1lri	98	1.45	1.2	S_9, Cl	17.3	Lascombe *et al.* (2002)
FERRICYTO	1ccr	112	1.50	22.8	S_4, Fe	8.7	Ochi *et al.* (1983)
PAZUR	1paz	123	1.55	1.1	S_6, Cu	16.3	Petratos *et al.* (1988)

In Table 2 we characterize the 18 structures which cannot be solved by *SIR*2004 at iteration zero (we use the NMRE algorithm only at iteration one) and those which are unsolved by standard *SIR*2004. For each structure we give the number of residues (Residues), the percentage of missing reflections below RES_{obs} (MIS), the heaviest-atom content and the *B* factor obtained by the Wilson plot.

Some details of the phasing process are given in Table 3 for the structures with $RES_{obs} > 1.2$ and in Table 4 for those with $RES_{obs} < 1.2$. The letter P or DM means that the starting phases of the trial leading to the correct solution were obtained by Patterson or by direct methods, respectively; the starting MPE value is given in parentheses. The values of $CORR_{exp}$ and $CORR_{extra}$ shown in the tables are calculated at the end of each DSR iteration; the value of fFOM shown in the tables is that calculated at the end of the phasing process (the process stops when fFOM > 1).

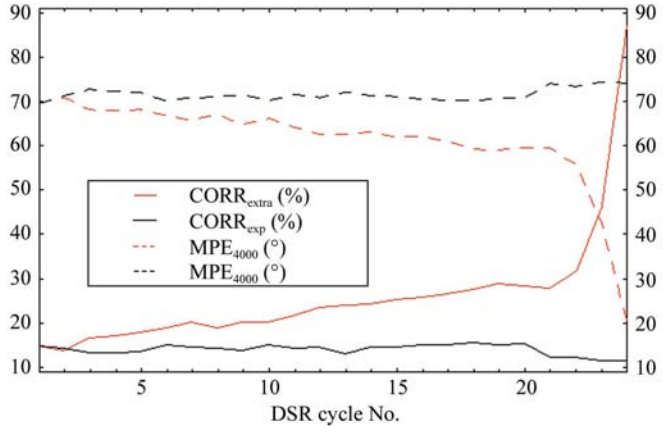

Figure 2
MB20: variations of $CORR_{extra}$, $CORR_{exp}$ (as percentages) and the MPE_{4000} of the 4000 stronger reflections as a function of the DSR cycle. The red and black curves correspond to the tests with and without the NMRE procedure, respectively.

We note the following.

(i) One (MB20) of the seven test structures with $RES_{obs} > 1.2$ Å (see Table 3) and five (ALPHA1, TETRAPLEX, DEOXY, METAXIA, SH2) of the ten structures with $RES_{obs} < 1.2$ Å (see Table 4) can only be solved by using the NMRE procedure.

(ii) Seven structures in Table 3 can be solved by both the procedures: on the whole they require 33 DSR iterations when the standard *SIR*2004 is used but only 27 when the NMRE algorithm is applied. Since the computing time necessary to complete a DSR iteration with the NMRE procedure is about 10% greater than the standard procedure, the use of the NMRE algorithm does not involve a longer computing time. The conclusion is that if $RES_{obs} > 1.2$ Å the supplementary information provided by the extrapolated intensities may accelerate the DSR process and in some cases can make the difference between success and failure.

(iii) Five structures in Table 4 can be solved by both procedures: on the whole they require 15 DSR iterations when the standard *SIR*2004 is used and 13 when the NMRE algorithm is applied. When the structure is solved by both the standard and the modified versions of *SIR*2004, usually (except for PROTG) $CORR_{exp} < CORR_{extra}$. This means that the use of the extrapolated reflections improves the quality of the final electron-density maps.

(iv) As is well known, incompleteness of the experimental data below RES_{obs} makes crystal structure solution more difficult, no matter which phasing method is used. Conversely, extrapolating such missed reflections together with the higher resolution reflections increases the amount of information carried on by the NMRE algorithm. It is probably not mere chance that for ALPHA1, CARBO, TETRAPLEX and SH2 the MIS values (see Table 2) are equal to 15, 9. 10.1 and 13%,

Table 3
Values of $CORR_{exp}$ and $CORR_{extra}$ for the subset of structures (among those listed in Table 2) with $RES_{obs} > 1.2$ Å.

See text for the meaning of the symbols.

Structure code	MPE (°)	$CORR_{exp}$	fFOM	$CORR_{extra}$	fFOM
CUPRE	P(67)	0.415; 0.821	1.4	0.415; 0.896	2.7
HEWL133	DM(68)	0.181; 0.229; 0.243; 0.330; 0.847	2.4	0.181; 0.230; 0.435; 0.899	2.9
WILD	DM(80)	0.337; 0.895	3.2	0.337; 0.916	4.1
ERBUTOX	P(73)	0.178; 0.213; 0.547; 0.732	1.3	0.178; 0.490; 0.758	1.5
MB20†	P(71)	0.150; 0.146; 0.155; 0.115	0.3	0.189; 0.235; 0.275; 0.872	3.6
COLE	DM(74)	0.222; 0.243; 0.237; 0.290; 0.372; 0.520; 0.779	1.2	0.222; 0.277; 0.474; 0.849	1.7
PAZUR	P(77)	0.592; 0.867	2.6	0.592; 0.897	3.5
FERRICYTO	P(71)	0.192; 0.222; 0.230; 0.227; 0.219; 0.247; 0.262; 0.310; 0.331; 0.329; 0.376; 0.403; 0.428; 0.434; 0.443; 0.477; 0.596; 0.737	2.0	0.192; 0.192; 0.236; 0.285; 0.316; 0.345; 0.357; 0.389; 0.433; 0.478; 0.502; 0.519; 0.556; 0.562; 0.596; 0.634; 0.665	2.2

† CORR values given per six iterations.

Table 4
Values of $CORR_{exp}$ and $CORR_{extra}$ for the subset of structures (among those listed in Table 2) with $RES_{obs} < 1.2$ Å.

See the text for the meaning of the symbols.

Structure code	MPE (°)	$CORR_{exp}$	fFOM	$CORR_{extra}$	fFOM
PYP	P(84)	0.122; 0.928	25.3	0.122; 0.151; 0.930	23.2
ALPHA1	DM(71)	0.160; 0.162; 0.062; 0.094; 0.085	0.0	0.160; 0.146; 0.113; 0.194; 0.947	7.7
TETRAPLEX†	P(71)	0.157; 0.148; 0.142; 0.149	−0.1	0.141; 0.159; 0.151; 0.802	3.5
GLPE	DM(74)	0.135; 0.193; 0.862	5.5	0.135; 0.194; 0.886	5.6
METAXIA	DM(74)	0.192; 0.199; 0.228; 0.215	0.0	0.192; 0.177; 0.242; 0.868	6.7
SH2	DM(74)	0.151; 0.143; 0.140; 0.152; 0.160	0.2	0.151; 0.170; 0.187; 0.253; 0.856	6.6
CALPO	DM(75)	0.162; 0.172; 0.217; 0.696; 0.919	8.3	0.162; 0.167; 0.184; 0.350; 0.922	9.3
PROTG	DM(78)	0.122; 0.910	7.4	0.122; 0.867	5.3
CARBO	P(76)	0.186; 0.206; 0.243; 0.258; 0.291; 0.332; 0.534; 0.871	4.3	0.186; 0.229; 0.266; 0.379; 0.913	4.3
DEOXY	P(77)	0.148; 0.141; 0.138; 0.145; 0.136; 0.148; 0.150; 0.143; 0.149	0.0	0.148; 0.167; 0.167; 0.188; 0.194; 0.203; 0.221; 0.229; 0.257; 0.317; 0.879	5.5

† CORR values given per six iterations.

respectively. However, extrapolation of only the missing reflections under RES_{obs} is not usually able to solve crystal structures that are unsolvable by standard *SIR*2004 (for brevity, we omit the experimental results).

(v) The number of DSR iterations for MB20, TETRAPLEX, FERRICYTO and DEOXY is quite high (23, 23, 16 and ten, respectively); evidently, in these cases the extrapolated values for both the intensities and the phases remain rough for a large number of iterations. To better understand the internal mechanism of NMRE it may be worthwhile to follow the variation of such estimates during the phasing process of MB20. In Fig. 2 we show for each DSR iteration the value of $CORR_{exp}$ and the mean phase error (MPE_{4000}) for the 4000 observed reflections with the largest value of $|E_{\mathbf{h}}^{obs}|$ when the standard version of *SIR*2004 is used (black lines). It may be noted that MPE_{4000} increases from one iteration to the next, while $CORR_{exp}$ decreases: the final map has a poorer information content than the starting one. In the same figure we show with red lines the values of $CORR_{extra}$ and of MPE_{4000} for the case in which the NMRE algorithm is used. $CORR_{extra}$ increases gradually cycle by cycle until the 21st DSR cycle, when $MPE_{4000} \simeq 60^{\circ}$; only a few iterations leads MPE_{4000} to 20° and $CORR_{extra}$ to 0.87. In Fig. 3 we give for each iteration the mean phase error (MPE_{3021}) for the 3021 extrapolated reflections actively used (10% of the reflections with resolution between RES_{obs} and RES_{ext}) and the corresponding residual

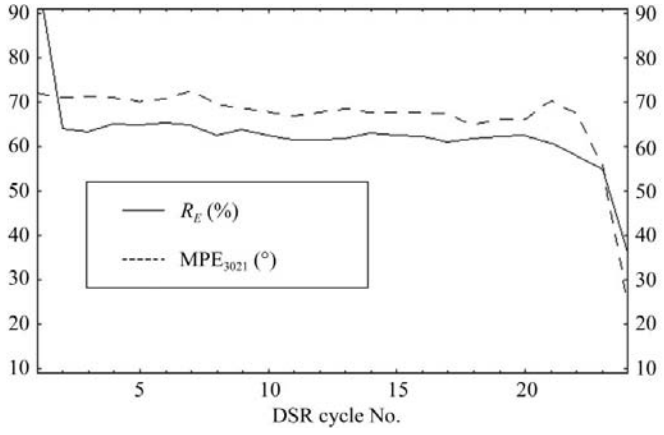

Figure 3
MB20: mean phase error (MPE_{3021}) for the 3021 extrapolated reflections actively used (10% of the reflections with resolution between RES_{obs} and RES_{ext}) and the corresponding residual R_E as a function of the DSR cycle.

$$R_E = \frac{\sum \big||E_{\rm true}| - |E_{\rm extra}|\big|}{\sum |E_{\rm true}|},$$

where $E_{\rm true}$ is calculated from the published structure. Both MPE_{3021} and R_E progressively decrease. The final values are 23.8° and 37.2%, respectively.

(vi) fFOM is a powerful figure of merit that correlates well with the quality of the electron-density map. For the seven structures in Table 3 solved by both the standard *SIR*2004 and by NMRE we find that ⟨fFOM⟩ is equal to 1.7 and 2.7, respectively. The analogous values for the five structures in Table 4 are 10.2 and 9.5, respectively [we have already noted on other occasions (Burla *et al.*, 2003) that fFOM values are usually larger for data at atomic resolution]. The conclusion is that the use of the NMRE algorithm improves the quality of the maps (by reducing the final MPE values) and for quasi-atomic resolution structures makes the recognition of the correct solutions among the different trials easier (because of the larger value of fFOM).

(vii) The trials given by the *SIR*2004 Patterson procedure very often lead to the correct solution. In our tests, we decided to spend more time on PM trials before exploring those obtained by DM. For example, for the nine test structures reported in Tables 3 and 4, labelled P in the MPE column, the Patterson trial which leads to the solution is always the first one, except for CUPRE and TETRAPLEX, for which success was obtained from the fourth and 17th Patterson trial, respectively. The CPU time needed to obtain an interpretable electron-density map from a Patterson trial for the structures in Tables 3 and 4 ranges from few tens of minutes (*e.g.* 22 min for PAZUR) up to several hours (*e.g.* 12.0 h for MB20) using a Xeon-1.7 GHz processor with a Linux operating system.

6. About the limitations of NMRE

Comparing Tables 2 and 4 suggests that NMRE is less useful when data resolution is atomic and MIS is simultaneously small. On the other hand, our tests indicate that when data resolution is worse than 1.2 Å and MIS is large, the NMRE procedure is maximally efficient, because it is able to retrieve supplementary information that is not experimentally available.

To check the limitations of the NMRE procedure as a function of the resolution, we have applied it to the observed data of PAZUR truncated at 1.6, 1.65 and 1.7 Å resolution; we extrapolated data to 1.2 Å in all three cases. The structure was solved both with data truncated at 1.6 Å (two iterations) and at 1.65 Å (six iterations); structure solution failed with data truncated at 1.7 Å (exploring up to 30 iterations). On the other hand, the standard *SIR*2004 succeeded only with data truncated to 1.6 Å. The conclusion is the following: when data resolution is too low (say, worse than about 1.6–1.7 Å also) in the presence of heavy atoms it is very difficult for NMRE to retrieve the missing experimental information and then to solve the structure.

Thus far our tests have checked the NMRE procedure by internal tests: *e.g.* by comparing the performances of *SIR*2004 with and without NMRE. To validate our results using a different package we applied *ACORN* (Foadi *et al.*, 2000), a well documented and powerful program included in *CCP*4 v.5.0 (Collaborative Computational Project, Number 4, 1994). *ACORN* is designed to solve protein structures starting from an initial phase set which (i) can be provided by the user, (ii) can be obtained by a random search procedure devoted to locate the potential heavy atoms or (iii) or can be obtained from the coordinates of a molecular fragment (even if small) suitably oriented and positioned by *ACORN* itself.

We first tried to phase *ab initio* the test structures that were unsolved by the standard *SIR*2004, but were solved if the NMRE procedure is applied (*e.g.* 1byz, 1nkd, 1d4t, 1a6n, 352d with $\mathrm{RES}_{\rm obs} < 1.2$ Å; 1dxd with $\mathrm{RES}_{\rm obs} > 1.2$ Å). *ACORN* applied the random search procedure to preliminarily locate the heavy atoms (sulfurs when no heavier atomic species are present; Yao *et al.*, 2002). As for the standard *SIR*2004, no solution was obtained.

We then entrusted to *ACORN* the task of extending and refining the initial phase sets resulting from the tangent or Patterson procedures implemented in *SIR*2004. The purpose was to check whether the powerful *ACORN* procedures were able to extend the phases and drive them to their correct values. Again, no solution was obtained for all the checked structures. This suggests that the quantity of information contained in the initial phase sets is not sufficient, at least for some of the best programs currently in use, for a successful phase extension and refinement, unless the supplementary information gained by the NMRE procedure is applied.

7. Conclusions

We have integrated into the *SIR*2004 program a novel procedure called NMRE which combined with classical electron-density modification techniques is able to (i) extrapolate moduli and phases of non-measured reflections with resolution lower or higher than the experimental one and (ii) actively use such moduli and phases in an *ab initio* phasing process. The procedure has been applied to experimental data of macromolecular crystal structures with resolution ranging from atomic to 1.5 Å.

The experimental applications show that the NMRE algorithm makes the phasing process more efficient, improves the quality of the electron-density maps and solves structures that are unsolvable *via* the standard *SIR*2004. Furthermore, the use of NMRE makes the recognition of the correct solution by means of our figure of merit fFOM easier.

These results may be related to the fact that the procedure of using few percent of an electron-density map in a typical Fourier inversion strengthens the positivity and the atomicity of the map and although one does not know with precision either the modulus or the phase of the unobserved reflections, their estimates have to be self-consistent with the imposed positivity and atomicity constraints. Such estimates in turn become a source of supplemental information that, if handled with care, can help in success in difficult cases, *e.g.* data sets that have proved resistant to other attempts.

APPENDIX *A*

We have used in this paper the approximation

$$ {}_1F_1\left(-\frac{1}{2};1;-z^2\right) \simeq \left(1+\frac{4z^2}{\pi}\right)^{1/2}, \qquad (5) $$

which is slightly different from that suggested by Burla, Carrozzini, Cascarano, Giacovazzo, Polidori *et al.* (2002),

$$ {}_1F_1\left(-\frac{1}{2};1;-z^2\right) \simeq \left(1+\frac{2z^2}{\pi^{1/2}}\right)^{1/2}. \qquad (6) $$

Both approximations have been obtained by studying the asymptotic behaviour and the power-series expansion of the function ${}_1F_1(-\frac{1}{2};1;-z^2)$.

Approximation (5) involves a maximum error of 4% for small R_p values (say $R_p < 0.4$), but the error is close to zero for higher R_p values. In contrast, approximation (6) leads to a maximum error of only 1% for small R_p values, but the error becomes about 5% for higher R_p values.

(5) has a notable property in our probabilistic context which is not shared by (6): when $\sigma_A \simeq 1$ (when the partial nearly coincides with the complete structure), it gives $\langle R|R_p\rangle = R_p$ and $\nu_1 = 0$, as common sense suggests.

References

Bacchi, A., Pelizzi, G., Sheldrick, G., Amari, G. M., Delcanale, M. & Redenti, E. (2002). *Supramol. Chem.* **14**, 67–74.

Banuelos, S., Saraste, M. & Carugo, K. D. (1998). *Structure*, **6**, 1419–1431.

Brunori, M., Vallone, B., Cutruzzola, F., Travaglini-Allocatelli, C., Berendzen, J., Chu, K., Sweet, R. M. & Schlichting, I. (2000). *Proc. Natl Acad. Sci. USA*, **97**, 2058–2063.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Spagna, R. (2005). *J. Appl. Cryst.* **38**, 381–388.

Burla, M. C., Carrozzini, B., Caliandro, R., Cascarano, G., De Caro, L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* A**59**, 560–568.

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

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2002). *Acta Cryst.* D**58**, 928–935.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Moustiakimov, M. & Siliqi, D. (2005). *Acta Cryst.* D**61**, 556–565.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005). *Acta Cryst.* A**61**, 343–349.

Chatani, E., Hayashi, R. & Moriyama, H. (2002). *Protein Sci.* **11**, 72–81.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

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

Derrick, J. P. & Wigley, D. B. (1994). *J. Mol. Biol.* **243**, 906–918.

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

Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137–1147.

Genick, U. K., Soltis, S. M., Kuhn, P., Canestrelli, I. L. & Getzoff, E. D. (1998). *Nature (London)*, **392**, 206–109.

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

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

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

Langs, D. A. (1998). *Acta Cryst.* A**54**, 44–48.

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

Lewis, H. A., Chen, H., Edo, C., Buckanovich, R. J., Yang, Y. Y. L., Musunuru, K., Zhong, R., Darnell, R. B. & Burley, S. K. (1999). *Structure*, **7**, 191–203.

Karle, J. & Hauptman, H. (1964). *Acta Cryst.* **17**, 392–296.

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

Lunin, V. Yu. & Urzhumtsev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.

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

Ochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, S. & Morita, Y. (1983). *J. Mol. Biol.* **166**, 407–418.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Petratos, K., Dauter, Z. & Wilson, K. S. (1988). *Acta Cryst.* B**44**, 628–636.

Phillips, K., Dauter, Z., Murchie, A. I., Lilley, D. M. & Luisi, B. (1997). *J. Mol. Biol.* **273**, 171–182.

Privé, G. G., Anderson, D. H., Wesson, L., Cascio, D. & Eisenberg, D. (1999). *Protein Sci.* **8**, 1400–1409.

Rappleye, J., Innus, M., Weeks, C. M. & Miller, R. (2002). *J. Appl. Cryst.* **35**, 374–376.

Seeman, N. C., Rosenberg, J. M., Suddath, F. L., Kim, J. J. & Rich, A. (1976). *J. Mol. Biol.* **104**, 109–144.

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 Academic Publishers.

Spallarossa, A., Donahue, J., Larson, T. & Bolognesi, M. (2001). *Structure*, **9**, 1117–1125.

Smith, J. L., Corfield, P. W. R., Hendrickson, W. A. & Low, B. W. (1988). *Acta Cryst.* A**44**, 357–368.

Srinivasan, R. & Ramachandran, G. N. (1965). *Acta Cryst.* **19**, 1008–1014.

Vaney, M. C., Maignan, S., Riès-Kautt, M. & Ducruix, A. (1996). *Acta Cryst.* D**52**, 505–517.

Vlassi, M., Dauter, Z., Wilson, K. S. & Kokkinidis, M. (1998). *Acta Cryst.* D**54**, 1245–1260.

Vojtechovsky, J., Chu, K., Berendzen, J., Sweet, R. M. & Schlichting, I. (1999). *Biophys. J.* **77**, 2153–2174.

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.

Yao, J. X., Woolfson, M., Wilson, K. S. & Dodson, E. J. (2002). *Z. Kristallogr.* **217**, 636–643.