Acta Cryst D
BIOLOGICAL
CRYSTALLOGRAPHY

ISSN 1399-0047

Advances in molecular-replacement procedures: the *REVAN* pipeline

Benedetta Carrozzini, Giovanni Luca Cascarano, Carmelo Giacovazzo* and Annamaria Mazzone

Istituto di Cristallografia, CNR, Via G. Amendola 122/o, 70126 Bari, Italy. *Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

Received 13 May 2015
Accepted 1 July 2015

Edited by Z. Dauter, Argonne National Laboratory, USA

Keywords: molecular replacement; sequence mutation; automated model building; *REVAN*.

The *REVAN* pipeline aiming at the solution of protein structures *via* molecular replacement (MR) has been assembled. It is the successor to *REVA*, a pipeline that is particularly efficient when the sequence identity (SI) between the target and the model is greater than 0.30. The *REVAN* and *REVA* procedures coincide when the SI is >0.30, but differ substantially in worse conditions. To treat these cases, *REVAN* combines a variety of programs and algorithms (*REMO*09, *REFMAC*, *DM*, *DSR*, *VLD*, *free lunch*, *Coot*, *Buccaneer* and *phenix.autobuild*). The MR model, suitably rotated and positioned, is first refined by a standard *REFMAC* refinement procedure, and the corresponding electron density is then submitted to cycles of *DM*–*VLD*–*REFMAC*. The next *REFMAC* applications exploit the better electron densities obtained at the end of the *VLD*–EDM sections (a procedure called vector refinement). In order to make the model more similar to the target, the model is submitted to mutations, in which *Coot* plays a basic role, and it is then cyclically resubmitted to *REFMAC*–EDM–*VLD* cycles. The phases thus obtained are submitted to *free lunch* and allow most of the test structures studied by DiMaio *et al.* [(2011), *Nature (London)*, **473**, 540–543] to be solved without using energy-guided programs.

1. Introduction

Several popular software packages are currently available for the X-ray crystal structure solution of macromolecules *via* molecular-replacement techniques (MR): some examples include *AMoRe* (Navaza, 1994), *MOLREP* (Vagin & Teplyakov, 2010), *Phaser* (McCoy *et al.*, 2007), *ULTIMA* (Rabinovich *et al.*, 1998), *REMO* and *REMO*09 (Caliandro *et al.*, 2006, 2009). Simultaneously, thanks to advances in automatic computing, six-dimensional space procedures have been developed, such as *EPMR* (Kissinger *et al.*, 1999), *Queen of Spades* (Glykos & Kokkinidis, 2000, 2004) and *SOMoRe* (Jamrog *et al.*, 2003).

To increase MR productivity, automated pipelines have been proposed such as *NORMA* (Delarue, 2008), *MrBUMP* (Keegan & Winn, 2008), *BALBES* (Long *et al.*, 2008), part of the JCSG software (Schwarzenbacher *et al.*, 2008) and automated servers such as *OCA* (Boutselakis *et al.*, 2003) and *PSI-BLAST* (Altschul *et al.*, 1997). For a given target, such software tries to perform the following steps.

(i) A list of search models (templates) potentially able to solve the target structure *via* MR techniques is found by a proper analysis of the Protein Data Bank (PDB). The templates are aligned against the target and selected on the basis of sequence identity (SI): templates with the largest values of SI are favoured. The MR process is expected to be straightforward if templates with an SI sufficiently larger than 0.30 are found; otherwise, it frequently fails.

(ii) The templates are modified to increase the signal-to-noise ratio by preserving the part of the model molecule which is in common to the target and pruning the part which is expected to lack correspondence. There are several ways to obtain such modified templates: polyalanine models, models with modified atomic thermal factors and models with the amino-acid sequence corrected in accordance with template–target alignment. More recently, techniques that incorporate local structural information as restraints have been introduced to further improve phase refinement (Schröder *et al.*, 2007, 2010; Headd *et al.*, 2012; Kidera & Gō, 1992; Delarue, 2008; Kleywegt & Jones, 1997; Cowtan, 1998; Terwilliger, 2001).

(iii) The top-ranked models are used by an MR program: when correctly located and translated, they are submitted to a refinement process and then to an automated model-building program such as *ARP/wARP* (Perrakis *et al.*, 1999), *phenix.autobuild* (Terwilliger *et al.*, 2008), *MAID* (Levitt, 2001), *MAIN* (Turk, 2013) or *Buccaneer* (Cowtan, 2006).

Two recent new approaches are strictly correlated with this paper. In DiMaio *et al.* (2011), algorithms for protein structure modelling are combined with those developed for crystal structure solution (see also Terwilliger *et al.*, 2012; Adams *et al.*, 2013). The scenario is the following. In difficult MR cases, when it is possible to correctly locate the model the resulting electron-density map is too noisy to be interpreted. This is usually the case when the SI is <0.30, where the atomic positions of the template and the target structures may differ by 2–3 Å. To overcome this problem, a suite using physically realistic all-atom potential functions, originally designed to predict protein structures given their amino-acid sequence (*i.e.* the program *Rosetta*; see Das & Baker, 2009), was used to identify the correct MR solution, when ambiguous, and to improve the model until it may be interpreted by combining force fields and experimental electron density.

In these difficult cases, the templates available at the end of steps (i) and (ii), suitably oriented and positioned by the chosen MR program, are remodelled during the phase expansion and refinement using supplementary information provided by energy-based procedures. The combination of *Rosetta* with density-modification techniques (EDM) and restrained reciprocal-space refinement led to the solution of MR cases for which more traditional approaches failed. The method extended the applicability of MR techniques to cases where the SI is close to 0.20.

The pipeline described by Carrozzini *et al.* (2013), here denoted *REVA* as a short reference (although this acronym was not used in the corresponding paper), is centred on the use of the *VLD* (*vive la difference*) algorithm, in combination for the first time with a non-*ab initio* (*e.g.* MR) phasing approach. The pipeline involves the use of *REMO09* (Caliandro *et al.*, 2009) as the MR program and *REFMAC* (Murshudov *et al.*, 2011) for reciprocal-space refinement, followed by *DM* (Cowtan, 1994), *DSR* (Giacovazzo & Siliqi, 1997), *VLD* (Burla *et al.*, 2011), *free lunch* (Caliandro *et al.*, 2005, 2007) and *ARP/wARP*. The pipeline automatically led to the solution of 40 of 45 test structures: for these, *ARP/wARP* automatically provided a sequence coverage of greater than 90%. However, only for eight of the test structures was the SI smaller than 0.50 and only for two was the SI close to 0.30.

While the first pipeline was mainly based on the use of energy-optimization algorithms and on structure rebuilding, *REVA* mainly benefits from the increased efficiency of EDM techniques when integrated with *VLD* and *free lunch*. Such methods, however, are hardly able to move significant parts of the templates closer to the target positions if they are at a distance of greater than 1.5–2 Å, and therefore do not allow *REVA* to efficiently process cases for which the SI is <0.30. It is therefore sensible to try to improve *REVA* by including supplementary techniques and auxiliary software capable of remodelling unaligned regions, optimizing the backbone, defining new side-chain torsion angles *etc.*, in such a way that the template gradually becomes more similar to the target during the phasing process. In this new pipeline, denoted *REVAN* for simplicity, no use will be made of energy-guided optimization techniques: their action, aiming at repositioning fragments of the model that are too distant from the target, is substituted by suitably combining *Coot* (Emsley *et al.*, 2010), *via Scheme* scripts, with *REFMAC*, EDM, *VLD* and *free lunch* techniques. The automated model-building (AMB) process is entrusted to *Buccaneer*, *phenix.autobuild* or *ARP/wARP*; *Buccaneer* is the AMB program that is automatically selected by the *REVAN* pipeline in the standard conditions; the user can modify the default choice.

We will also show that our approach is substantially different from the morphing techniques originally formulated by Terwilliger *et al.* (2012) to modify and relocate, without any use of energy-based programs, models that are not sufficiently close to the target structure.

The *REVAN* algorithms are described in §2: in §2.1 the figures of merit used in most of the steps of our phasing procedure and designed to automatize the full phasing process are summarized. The experimental results are illustrated in §3.

2. The new pipeline architecture and its methods

In §2.1 the figures of merit of *REVAN* are described, while the architecture of the program and methods are described in §2.2.

2.1. Figures of merit

In the new pipeline architecture (see §3), three figures of merit are used to allow the program to take sensible decisions and therefore to automate the phasing process.

(i) $R_{\rm cryst}$, where

$$R_{\rm cryst} = \frac{\sum_{\mathbf{h}} |F_{\rm obs}| - |F_{\rm calc}|}{\sum_{\mathbf{h}} |F_{\rm obs}|}.$$

(ii) $R_{\rm free}$ (Brünger, 1992), as calculated by *REFMAC* or by the AMB program.

(iii) fFOM2, defined by

$${\rm fFOM2} = {\rm fFOM} \cdot [{\rm CC(all)}_{\rm current}]^{1/2},$$

where

$$\mathrm{fFOM} = \frac{\mathrm{RAT}_{\mathrm{current}}}{\mathrm{RAT}_{\mathrm{initial}}}\frac{\mathrm{CC(all)}_{\mathrm{current}}}{\mathrm{CC(all)}_{\mathrm{initial}}}\frac{\mathrm{CC(large)}_{\mathrm{current}}}{\mathrm{CC(large)}_{\mathrm{initial}}}.$$

RAT = $\mathrm{CC}_{w,R}/\langle R^2_{\mathrm{calc}}\rangle_{\mathrm{weak}}$, where R_{calc} are the amplitudes of the normalized structure factors obtained by inversion of the current electron-density map, and the average $\langle R^2_{\mathrm{calc}}\rangle_{\mathrm{weak}}$ is calculated over 30% of the measured reflections (those with the weakest $|F_{\mathrm{obs}}|$ values). $\mathrm{CC}_{w,R}$ is the correlation coefficient between the largest R_{obs} amplitudes (about 70% of the total) and the corresponding weights.

CC is the correlation factor between R_{obs} and R_{calc}; the words 'all', 'large' and 'weak' indicate the overall set of normalized structure factors, the subset (70%) of the largest $|F_{\mathrm{obs}}|$ values and the subset (30%) of the weakest ones, respectively.

While fFOM estimates the relative phase improvement (from the 'initial' to the 'current' state), fFOM2 includes an absolute estimate of the quality of the phases since it involves the current value of CC.

2.2. Architecture of the program and algorithms

The algorithms described in this paper have the main purpose of automatically solving, *via* MR and ancillary techniques, crystal structures with an SI of <0.30. We will apply them to the severe test constituted by structures resistant to *REVA* and the structures used by DiMaio *et al.* (2011) (see Fig. 1 in the Supplementary Information for that paper). The phasing approach of DiMaio and coworkers may be schematized as follows.

(i) Templates were identified using *HHpred* (Söding, 2005), which was also used to generate the initial alignment. Templates were prepared by removing unaligned residues and by stripping non-identical side chains to the γ C atom.

(ii) The MR solutions were obtained by *Phaser* and were then submitted to *Rosetta* to rebuild gaps in the initial alignment and in regions around deleted residues.

(iii) The application of energy refinement, restrained by the electron-density maps, led to new models, the best of which were again submitted to an iterative rebuild with *Rosetta* in combination with *phenix.autobuild*.

(iv) If the final model builds the majority of the protein with an R_{free} of <0.4, the structure was considered to be solved.

The above procedure is very efficient when looking at the results: it was able to solve, amongst others, the 13 structures reported in Table 1, all characterized by an SI of <0.30. The first five of these could be solved by combining *phenix.autobuild* with simulated annealing in Cartesian or in torsion space, or with deformable elastic network refinement (DEN; Schröder *et al.*, 2010). Structures 6–13 were resistant to any solution attempt and were only solved by combining *phenix.autobuild* with *Rosetta*: such a combination was also able to solve the first five test structures.

The procedure, however, requires considerable computing time: indeed, up to several thousand *Rosetta* models should be generated for each structure, and this leads to an overall CPU time that varies from approximately 30 to 130 h per structure.

Table 1
Basic parameters for DiMaio *et al.* test structures.

No. is the number of the target structure as ordered in DiMaio *et al.* (2011), $\mathrm{PDB_T}$ is the PDB code of the target structure, RES is the experimental data resolution in Å and SI is the sequence identity with the model. R_{free} is the final value obtained by DiMaio and coworkers.

No.	$\mathrm{PDB_T}$	RES	SI (%)	R_{free}
1	3o8s	2.12	22	0.31
2	3nng	2.18	19	0.29
3	4fqd	2.54	27	0.27
4	3npg	2.7	21	0.30
5	—	—	—	—
6	3nr6	1.96	30	0.34
7	3q60	2.05	22	0.28
8	—	—	—	—
9	—	—	—	—
10	3tx8	3.17	20	0.39
11	3zyt	2.45	18	0.27
12	4e2t	1.75	100	0.29
13	3ons	2.9	29	0.39

REVA is unable to solve the 13 test structures of DiMiao and coworkers both because of the difficulties encountered by *REMO*09 and because of the subsequent unsuccessful phase-refinement step. The aim of this paper is to show that refinement of the MR model may be obtained without any use of energy-guided approaches, and therefore that the corresponding computing time may be considerably reduced. Accordingly, in the new procedure *REVAN*, the heir to *REVA*, we will use the same templates and, when possible, the same MR solutions as used by DiMaio and coworkers (the authors have kindly made most of the experimental data and intermediate results available at http://www.phenix-online.org/phenix_data/terwilliger/rosetta_2011/).

The main steps of *REVAN* may be described as follows.

Step 1. Reading the MR solution. The procedure starts with the MR solution. Let us call the electron density and the phases obtained at the end of this step ρ_{MR} and φ_{MR}, respectively, and let $\langle|\Delta\varphi_{\mathrm{MR}}|\rangle$ represent the corresponding average phase error. When the template is poor, as in all of our test cases, ρ_{MR} will be very noisy and φ_{MR} will be far from the corresponding published phase value. In the following, $\mathrm{M_{MR}}$ is the model available after this MR step.

Step 2. Template alignment. The selected template is aligned with the target structure by using a modified version of the Needleman & Wunsch (1970) dynamic alignment algorithm and the SI parameter is estimated. The terminal loop residues are cut.

Step 3. REFMAC model refinement. *REFMAC* is applied to the model structure in default conditions. The number of cycles is guided by R_{cryst}: the summations are over all of the measured reflections and F_{calc} is the structure factor calculated from the current model under refinement. The number of *REFMAC* cycles is defined by the following criterion. R_{cryst} is calculated every 15 cycles. At cycle $15(n+1)$ R_{cryst} is compared with the value obtained at cycle $15n$. If it is larger, and if the average absolute difference between the phase values at cycle

$15n$ and the phase values at cycle $15(n-1)$ is less than 3°, then the program stops and the phases obtained at cycle $15n$ are restored. It is useful to notice that the above criterion often allows a large number of *REFMAC* cycles, in general between 75 and 150: owing to the robustness of the *REFMAC* algorithms this overwork often leads to significant model improvements.

The last *REFMAC* cycle provides a new model template, $\mathrm{M_{RF}}$, defined by a set of atomic positions and by the corresponding vibrational factors. In the following, the phases and the electron density corresponding to the final *REFMAC* model will be denoted φ_{RF} and ρ_{RF}, respectively; $\langle|\Delta\varphi_{\mathrm{RF}}|\rangle$ represents the corresponding average phase error.

The sequence EDM–*VLD*–EDM is applied to refine the electron-density map calculated using the $\mathrm{M_{RF}}$ template.

Step 4. REFMAC application with phased option (REFMAC$_{P}$*) and electron-density modification*. The sequence *REFMAC*$_{\mathrm{P}}$–EDM–*VLD*–EDM is automatically launched six times. The EDM routines of Cowtan (1994) are used for the EDM step; when, according to internal criteria, EDM cycles stop then the difference electron density is calculated according to the *VLD* algorithm and is combined with the previous model electron density to provide the set of new model phases. Additional EDM cycles refine the phases produced by the *VLD* step. Let φ_{VE} represent the target phase estimate available at the end of each *VLD*–EDM cycle, $\langle|\Delta\varphi_{\mathrm{VE}}|\rangle$ be the corresponding average phase error and ρ_{VE} indicate the resulting electron density as obtained by Fourier inversion of φ_{VE}.

Just after the application of the first EDM–*VLD*–EDM sequence two sources of information are available: the last *REFMAC* model $\mathrm{M_{RF}}$ obtained in Step 3, with the corresponding phases φ_{RF}, and the phases φ_{VE} just obtained, with the corresponding electron density ρ_{VE}. Since the φ_{VE} values usually estimate the true phases better than φ_{RF}, a modified structure refinement is performed which has its roots in the work of Arnold & Rossmann (1988) and is denoted vector refinement. They observed that when exceedingly good phase information is available, as may occur in virus crystallography where noncrystallographic symmetry may be exploited, such information may be used to improve molecular-replacement phases.

Our situation does not fit the Arnold and Rossmann conditions: indeed, the average phase error for the φ_{VE} set is only a few degrees smaller than the error corresponding to the φ_{VE} set. However, we verified that a such an algorithm may be usefully applied. Giacovazzo (2015) showed that such refinement minimizes the difference between the current electron density, as computable directly from the *REFMAC* model (in our case ρ_{RF}), and the electron density corresponding to the higher quality reflections (in our case ρ_{VE}).

To improve $\mathrm{M_{RF}}$, just after the application of each EDM–*VLD*–EDM sequence we use a restrained vector refinement implemented in *REFMAC* (the option denoted as phased maximum likelihood). Such special refinement corresponds, in direct space, to adapting $\mathrm{M_{RF}}$ to the electron density ρ_{VE} without passing through the *ex novo* model-rebuilding step, which may fail because ρ_{RF}, at this stage, is still of poor quality.

The final result of this step is represented by φ_{RFP4}: the subscript RFP indicates that the target phase estimates were obtained by the last cycle of *REFMAC*$_{\mathrm{P}}$, the phased maximum-likelihood version of *REFMAC*, and the number 4 indicates that the phases are obtained at the end of Step 4. $\langle|\Delta\varphi_{\mathrm{RFP4}}|\rangle$ is the corresponding average phase error, $\varphi_{(\mathrm{VE})4}$ and $\langle|\Delta\varphi_{(\mathrm{VE})4}|\rangle$ are the target phase estimate available at the end of the last EDM cycle and the corresponding average phase error, respectively, and $\rho_{(\mathrm{VE})4}$ denotes the corresponding electron-density map.

At the end of Step 4, the figures of merit R_{cryst4}, R_{free4} (as computed by *REFMAC*$_{\mathrm{P}}$) and $\mathrm{fFOM2_4}$ are calculated. If R_{free4} is significantly smaller than 0.4 then the structure may be considered to be solved and the model is submitted to the final AMB process (as for structure 12 in Table 1) to save CPU time.

Step 5. Polyalanine remodelling and electron-density modification. If Steps 1–4 are unable to solve the target structure, the template is remodelled in order to be more similar to the target. Firstly, a polyalanine model is created and cycles of *REFMAC*$_{\mathrm{P}}$ are used to refine this pruned model: $\rho_{(\mathrm{VE})4}$, as obtained at the end of Step 4, is the reference electron density for this application.

The EDM–*VLD*–EDM sequence is launched to obtain a new and possibly better electron density $\rho_{(\mathrm{VE})5}$.

As a result, $\langle|\Delta\varphi_{\mathrm{RFP5}}|\rangle$ and $\langle|\Delta\varphi_{(\mathrm{VE})5}|\rangle$ are the average phase errors calculated after *REFMAC*$_{\mathrm{P}}$ and at the end of the last EDM cycle, respectively. At the end of Step 5, R_{cryst5} and $\mathrm{fFOM2_5}$ are calculated and are actively used as stopping criteria: if $\mathrm{fFOM2_5} < \mathrm{fFOM2_4}$ or if $(R_{\mathrm{cryst5}} - R_{\mathrm{cryst4}}) > 0.01$, the *REVAN* procedure restores the template (and the corresponding phase set) obtained at the end of Step 4 and directly applies Step 8.

Step 6. Mutation of model residues. According to the alignment in Step 2, *Coot* is applied to automatically mutate model residues with low thermal factors. The rationale for this choice is the following: low thermal factors are expected to characterize the C^{α} atoms that are more carefully located. It should then be more easy to find satisfactory positions for the mutated residues.

The best rotamer orientations, selected using the *MolProbity* library (Lovell *et al.*, 2000), are scored by searching for the best fit between $\mathrm{M_{RF}}$ and the current electron density ρ_{VE}. The basic conditions for success are the following: the selected C^{α} atoms are located by *REFMAC* with sufficient accuracy, and the reference electron-density map ρ_{VE} is a good guide for positioning the new residues. The above conditions are not always well satisfied in the cases considered in this paper, but it is supposed that *Coot* may work well if the phase error originally obtained at the end of the MR step has been reduced during Steps 2–6. Accordingly, at the end of this step some protein-chain fragments of variable length docked into the protein sequence and almost correctly placed may be available. Such a model is submitted to *REFMAC*$_{\mathrm{P}}$ refinement

Table 2
Average phase errors for DiMaio *et al.* test structures at different steps of the phasing procedure.

For each of the test structures from DiMaio *et al.* (2011) considered in this paper we show the following. (i) The structure number as reported in Table 1 (No.). (ii) The average phase errors obtained at the end of Step 2 (column 2). (iii) The average phase error obtained at the end of the various steps by *REFMAC* (standard or vector refinement mode) and, on a second line, the average phase error after the last EDM cycle (columns 3–8). n.p. indicates that related step of the *REVAN* procedure was not performed (see §3 for details). (iv) The final average phase error ($\langle|\Delta\varphi_{\rm FIN}|\rangle$) obtained after application of the automated model-building program (using *Buccaneer* as the default; values in bold were obtained using *phenix.autobuild*) and the corresponding $R_{\rm free}$ value (last column). All phase errors are in degrees.

		$\langle\lvert\Delta\varphi_{\rm RF}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP4}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP5}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP6}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP7}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP8}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm FIN}\rvert\rangle$
No.	$\langle\lvert\Delta\varphi_{\rm MR}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)4}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)5}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)6}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)7}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)8}\rvert\rangle$	$R_{\rm free}$
1	76	65	66	61	63	62	60	31
		57	58	54	54	54	53	0.29
2	76	70	66	64	66	67	n.p.	**62**
		67	63	61	63	64	n.p.	**0.48**
3	75	69	66	56	58	n.p.	55	27
		65	63	54	54	n.p.	53	0.29
6	81	53	52	51	n.p.	n.p.	51	30
		48	46	46	n.p.	n.p.	46	0.29
7	81	60	59	58	58	60	n.p.	30
		57	55	54	55	55	n.p.	0.30
10	75	51	51	51	n.p.	n.p.	51	**47**
		47	47	46	n.p.	n.p.	46	**0.44**
11	84	72	67	68	70	72	n.p.	27
		66	59	58	57	59	n.p.	0.29
12	75	35	34	n.p.	n.p.	n.p.	n.p.	25
		32	31	n.p.	n.p.	n.p.	n.p.	0.29
13	73	50	50	50	49	50	49	32
		44	44	44	43	43	43	0.32

using the electron-density map obtained at the end of Step 6 as a reference.

As a result, $\langle|\Delta\varphi_{\rm RFP6}|\rangle$ and $\langle|\Delta\varphi_{\rm (VE)6}|\rangle$ are the average phase errors calculated after *REFMAC*$_{\rm P}$ and at the end of the last EDM cycle, respectively. The figures of merit are computed even at this stage and actively used once more: if $\text{fFOM2}_6 < \text{fFOM2}_5$ or if $(R_{\rm cryst6} - R_{\rm cryst5}) > 0.01$, *REVAN* restores the model (and the phases) obtained at Step 5 and goes to Step 8.

Step 7. Filling gaps. Gaps between the protein-chain fragments are filled using *Coot*: the fragments are extended *via* additional residues at the N- and C-termini according to the protein-model alignment. Such extended protein-chain fragments are then submitted to *REFMAC*: at the end a new model $M_{\rm RF}$ is available, the phases $\varphi_{\rm RF}$ of which are used to calculate a new electron-density map, which is then submitted to cycles of EDM–*VLD*–EDM–*REFMAC*$_{\rm P}$. At the end of this step, a new electron density $\rho_{\rm FL}$ is available.

As a result, $\langle|\Delta\varphi_{\rm RFP7}|\rangle$ and $\langle|\Delta\varphi_{\rm (VE)7}|\rangle$ are the average phase errors calculated after *REFMAC*$_{\rm P}$ and at the end of the last EDM cycle, respectively. $R_{\rm cryst7}$ and fFOM2$_7$ are calculated: if $\text{fFOM2}_7 < \text{fFOM2}_6$ or if $(R_{\rm cryst7} - R_{\rm cryst6}) > 0.01$, the procedure restores the model (and the phases) of Step 6.

Step 8. REFMAC application with phased option and electron-density modification. The sequence *REFMAC*$_{\rm P}$–EDM–*VLD*–EDM is launched up to five times.

As a result, $\langle|\Delta\varphi_{\rm RFP8}|\rangle$ and $\langle|\Delta\varphi_{\rm (VE)8}|\rangle$ are the average phase errors calculated after *REFMAC*$_{\rm P}$ and at the end of the last EDM cycle, respectively. At the end of any sequence, the figures of merit are calculated again and used to stop the cyclic procedure and save CPU time.

Step 9. Extrapolation of unobserved reflections and automatic model building. *free lunch* is launched: the structure-factor extrapolation is used to further reduce the average phase error and therefore to make automated model building easier. *Buccaneer* automatically starts to perform the AMB process. The program stops if success is obtained (in practice, if $R_{\rm free}$ is significantly smaller than 0.4), otherwise the phases obtained from the *Buccaneer* model are used as the starting point for an additional application of the EDM–*VLD*–EDM sequence. The resulting $\varphi_{\rm VE}$ values are then used as a starting point for a new AMB application.

The EDM–*VLD*–EDM–*free lunch*–*Buccaneer* sequence is cycled up to ten times, and stops if $R_{\rm free}$, as calculated by *Buccaneer* at cycle $(n+1)$, is larger than the previous value or is stationary.

The *REVAN* algorithms described above may be usefully compared with those used by the morphing techniques described by Terwilliger *et al.* (2012). Both approaches require a starting model and a starting electron density. The morphing approach needs to identify, for each residue of the model, a proper translation which is applied to all of the atoms close to the residue. The translation for each residue is found by the *FFT* algorithm (Cowtan, 1998) and corresponds to the best fit between the atoms near to the C$^{\alpha}$ atom of each residue and the target electron-density map. The best translation vectors are then smoothed to take into account the possible variation of the shifts along a chain.

In *REVAN* no shift per residue is searched. The morphing approach is replaced by a different principle: the improvement of the model is based on the capacity of obtaining (*via* the *VLD* and *free lunch* techniques) an electron density better than the model refined by *REFMAC*. Since this feature is maintained in all of the *REVAN* steps, the current *REFMAC* model is adapted to each new density map *via* the restrained vector refinement mode and thus is substantially improved.

3. Applications

REVAN has been applied to the following set of test structures: (i) the abovementioned test structures employed by DiMaio *et al.* (2011) and quoted in Table 1, to show that *REVAN* may also succeed when the SI is <0.30, even without the help of energy-based programs, and (ii) the same set of structures employed to verify the efficiency of the *REVA* pipeline, constituted by four structures with high-resolution data (conventionally, better than 1.25 Å) and 41 structures with lower resolution data (from 1.50 to 2.86 Å). Some structures were not originally solved by MR, while others were

Table 3
REVAN results for five of the test structures used for checking *REVA* (Carrozzini *et al.*, 2013).

The first column gives the PDB code; the headings for the other columns are the same as those in Table 2.

		$\langle\lvert\Delta\varphi_{\rm RF}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP4}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP5}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP6}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP7}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm RFP8}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm FIN}\rvert\rangle$
PDB code	$\langle\lvert\Delta\varphi_{\rm MR}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)4}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)5}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)6}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)7}\rvert\rangle$	$\langle\lvert\Delta\varphi_{\rm (VE)8}\rvert\rangle$	$R_{\rm free}$
1cgn	73	56	53	49	44	45	n.p.	23
		47	44	41	37	38	n.p.	0.20
1cgo	74	69	68	63	46	44	44	28
		62	62	55	40	38	38	0.30
1lat	71	59	60	59	61	n.p.	61	50
		56	55	55	58	n.p.	58	0.41
2iff	62	66	69	n.p.	81	n.p.	81	87
		76	72	n.p.	83	n.p.	83	0.49
2pby	79	45	44	42	39	36	36	27
		40	40	39	37	35	34	0.26

used as test cases by three-dimensional or six-dimensional MR search programs. For details, the reader is referred to Carrozzini *et al.* (2013).

Let us first deal with the test structures of DiMaio and coworkers. As specified in §3, the same templates and, when available, the same MR solutions as DiMaio *et al.* (2011) were used. In Table 1 we quote, in the original order (given by number), the PDB codes of each test structure, the data resolution (RES, in Å), the SI value (as a percentage) and the final $R_{\rm free}$ obtained after the *Rosetta* + *phenix.autobuild* approach (the penultimate column of Table 1 in DiMaio *et al.*, 2011). We notice the following.

(i) The data for structures 5 and 8 were not deposited by DiMaio and coworkers. Therefore, these structures were not used as objects of this study.

(ii) Structure 9 is a model arising from the combination of many PDB models. Our program code is not able to deal with this and therefore this structure was not included in our tests.

(iii) For structures 1, 2, 3, 6 and 11 the correctly positioned molecular models were not deposited: we positioned them *via Coot* and used these models as the starting points for our procedure.

As stated in §1, *REVA* is unable to solve the test structures of DiMaio and coworkers even when applied to the MR solutions provided by the authors. Indeed, *REVA* does not exploit either *REFMAC* vector refinement or the mutation algorithm (including *Coot*) described in §3.

REVAN is much more effective than *REVA*. Steps 1–4 of its procedure have a specific function: they make the molecular model as close as possible to the target structure before introducing the sequence mutations. In Steps 5–8 the approach is combined with sequence mutation to simplify structure recovery. In order to understand how success may be obtained, let us follow the trend of the phase error in steps 2–8 in Table 2: for each step we quote the phase error calculated by the last *REFMAC* cycle and, below, that at the end of the last EDM cycle of that step.

We observe the following.

(i) Usually, the first *REFMAC* refinement considerably reduces the average phase error obtained at the end of the MR step: compare the $\langle|\Delta\varphi_{\rm MR}|\rangle$ column with the $\langle|\Delta\varphi_{\rm RF}|\rangle$ column. The number of cycles is automatically fixed by the procedure (see Step 3 in §3) and in rare cases may also exceed 150. At a first sight the number of *REFMAC* cycles used may appear to be rather large, but such intensive use is often quite useful for the success of *REVAN*.

(ii) At the end of Step 4 the average phase error corresponding to the last cycle of *REFMAC* in vector refinement mode is usually significantly smaller than the mean phase error at the end of the MR step. In symbols, $\langle|\Delta\varphi_{\rm RFP4}|\rangle < \langle|\Delta\varphi_{\rm MR}|\rangle$. Since both $\langle|\Delta\varphi_{\rm RFP4}|\rangle$ and $\langle|\Delta\varphi_{\rm MR}|\rangle$ correspond to molecular models [which is not the case for $\langle|\Delta\varphi_{\rm RF}|\rangle$ and $\langle|\Delta\varphi_{\rm (VE)4}|\rangle$], both of the models might be submitted to a mutation process, but using the phases obtained at Step 4 is expected to be much more safe owing to the fact that the success of mutation mainly depends on the quality of the current electron-density map. This is also true for the molecular models refined at higher order steps.

(iii) It is always the case that $\langle|\Delta\varphi_{{\rm (VE)}i}|\rangle < \langle|\Delta\varphi_{{\rm RFP}i}|\rangle$ for $i = 4, \ldots, 7$. The smaller phase error corresponding to the electron-density map refined by EDM–*VLD*–EDM cycles gradually allows *REFMAC*, used in vector refinement mode, to virtuously distort the template and to make it more similar to the target.

(iv) The adopted figures of merit described in §§2 and 3 allow the automatic procedure to make the necessary decisions. Those shown as n.p. in Table 2 correspond to rejected mutations. As a consequence, the phase determined at the end of the last accepted step is used to start the automated model-building process.

(v) The average phase error of structure 2 has been improved by the *REVAN* pipeline (from 76 to 62°), but not sufficiently to succeed. Also, structure 4 remained unsolved (not shown in the table).

(vi) All eight of the other test structures were solved by *REVAN*. For seven of them the default procedure automatically leads to a very small final phase error $\langle|\Delta\varphi_{\rm FIN}|\rangle$ (obtained at the end of the model-building step) and the corresponding $R_{\rm free}$ values are significantly smaller than 0.40. They are fully comparable with the values corresponding to the final $R_{\rm free}$ value obtained by DiMaio and coworkers and reported in Table 1 and with the values obtained by Terwilliger *et al.* (2012) using the morphing approach (see Table 3 in that paper).

(vii) For structure 10, the *REVAN* procedure makes good progress. The average phase error corresponding to the MR step (75°) is reduced to 46° just before the automated model-building application. However, *Buccaneer* fails and *phenix.autobuild* partially succeeds (probably because of the poor data resolution of 3.17 Å), ending with 47° average error. The value of $R_{\rm free}$ clearly shows this partial success. Since the average phase error is sufficiently small, it is expected that direct inspection of the map may lead to a satisfactory structural model.

Let us now consider the 45 structures used by Carrozzini *et al.* (2013) to test *REVA*. This procedure is less time-consuming than *REVAN*. Therefore, the following default approach has been chosen for *REVAN*: if the SI is >0.35 then *REVA* is applied. If the structure is not solved by *REVA*, then *REVAN* is automatically applied. Accordingly, *REVAN* is applied as first choice only if the SI is <0.35.

Among the set of 45 test structures mentioned above only two, 1cgo and 1cgn, have an SI of <0.35 (0.30 and 0.31, respectively). Furthermore, only three of the 45 test cases with an SI of >0.35 remained unsolved by *REVA* in default mode: 1lat, 2pby and 2iff.

We applied *REVAN* to the above five structures using the same data as used for applying *REVA* with starting phases provided by *REMO09* and molecular models 2ccy, 2ccy, 1glu, 1mki and 1hem for 1cgo, 1cgn, 1lat, 2pby and 2iff, respectively.

The results are shown in Table 3. For three of the five structures (1cgn, 1cgo and 2pby) a satisfactory model is found by *Buccaneer*; for 1lat *phenix.autobuild* provided a more approximate model. 2iff resisted *REVAN*, probably because the scattering power of the MR model is too small a percentage (about 0.23) of the target.

After the submission of this paper, one of the referees suggested combining the method described above with the jelly-body approach recently introduced in *REFMAC*5 (Murshudov *et al.*, 2011). The results are described and are commented on in Appendix *A*.

4. Conclusions

The *REVAN* pipeline, addressed at solving crystal structures *via* MR, has been described and applied to difficult cases. In particular, the set of crystal structures used by DiMaio and coworkers to demonstrate the usefulness of combining MR procedures with energy-guided programs (*i.e. Rosetta*) has been considered. All such structures were characterized by very small values of sequence identity (less than 0.30) to the corresponding MR models, and therefore were resistant to usual MR approaches.

We showed that *REVAN* is able to solve most of the above test structures in an automatic way. A basic condition for success is the following. The use of the sequence EDM–*VLD*–EDM, applied to phases corresponding to a *REFMAC* molecular model, provides electron density with an average phase error significantly better than that corresponding to the *REFMAC* model. *REFMAC*, with the phased maximum-likelihood option, may then accommodate a better molecular model in this density. The procedure is cyclic.

Once the average phase error has been diminished, *Coot* may mutate the residues according to the sequence alignment between the model and target. Further cycles of EDM–*VLD*–EDM–*free lunch* may further improve the phase estimates.

REVAN also succeeded with the set of test structures used for checking the usefulness of *REVA*, particularly for the subset that was resistant to *REVA*.

Table 4
Results obtained at the end of the *PROCJB* procedure for the test structures of DiMaio and coworkers and for the test structures quoted in Table 3.

Structure code	$\langle|\Delta\varphi_{\rm RFP8}|\rangle$ $\langle|\Delta\varphi_{\rm (VE)8}|\rangle$
1	80 76
2	62 59
3	54 53
6	50 46
7	52 49
10	52 46
11	63 53
12	36 33
13	50 44
1cgn	46 39
1cgo	55 47
1lat	59 54
2iff	78 79
2pby	41 37

The important practical aspect of this work is that MR cases in which the SI is very low might be solved without using energy-guided programs: the crystal structure solution may be more friendly and a large amount of computing time may be saved. The purpose is similar to the morphing approach described by Terwilliger *et al.* (2012), but is achieved *via* quite different algorithms.

APPENDIX *A*

The suggestion by a referee to combine our algorithms with the jelly-body option of *REFMAC*5 has been tested by allowing its automatic use *via* suitable scripts. Such a combination implies that *REFMAC* models are now improved by coupling the restraints provided by the electron densities obtained *via*EDM–*VLD*–EDM techniques with the typical restraint of the jelly-body option, *i.e.* the regularization function

$$\sum_{d_{ij,{\rm current}}<d_{\max}} w(d_{ij}-d_{ij,{\rm current}}) \qquad (1)$$

for all pairs (i, j) belonging to the same chain, under the default limitation of 4.25 Å for the maximum distance (*i.e.* under control). Let us refer to this procedure as *PROCJB*. Its results are summarized in Table 4, where for each of the structures quoted in Tables 2 and 3 we report the values $\langle|\Delta\varphi_{\rm RFP8}|\rangle$ and $\langle|\Delta\varphi_{\rm (VE)8}|\rangle$. We notice the following,

(i) No substantial difference may be found from our standard procedure for structures 3, 6, 10, 12, 13, 1cgn and 2iff. In practice, the additional use of the function (1) does not lead to better molecular models.

(ii) *PROCJB* improves the models provided by our standard approach for structures 2, 7, 11 and 1cgo. In some way, the jelly-body application anticipates the model regularization performed in the subsequent step by the automatic model-building programs.

(iii) Structure 1 unexpectedly remains unsolved if *PROCJB* is employed. It is very likely that the supplementary use of the restraints (1) makes the refinement more rigid and does not allows effective refinement.

It may be concluded that the supplementary use of the jelly-body option of *REFMAC5* may be combined with the phasing procedure described in this paper with beneficial effects on the quality of the model, but its effective use requires some further study to avoid excessive rigidity in the model refinement.

References

Adams, P. D., Baker, D., Brunger, A. T., Das, R., DiMaio, F., Read, R. J., Richardson, D. C., Richardson, J. S. & Terwilliger, T. C. (2013). *Annu. Rev. Biophys.* **42**, 265–287.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.

Arnold, E. & Rossmann, M. G. (1988). *Acta Cryst.* A**44**, 270–283.

Boutselakis, H. *et al.* (2003). *Nucleic Acids Res.* **31**, 458–462.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2011). *J. Appl. Cryst.* **44**, 1143–1151.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Mazzone, A. M. & Siliqi, D. (2006). *J. Appl. Cryst.* **39**, 185–193.

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

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2007). *J. Appl. Cryst.* **40**, 931–937.

Caliandro, R., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2009). *Acta Cryst.* A**65**, 512–527.

Carrozzini, B., Cascarano, G. L., Comunale, G., Giacovazzo, C. & Mazzone, A. (2013). *Acta Cryst.* D**69**, 1038–1044.

Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.

Cowtan, K. (1998). *Acta Cryst.* D**54**, 750–756.

Cowtan, K. (2006). *Acta Cryst.* D**62**, 1002–1011.

Das, R. & Baker, D. (2009). *Acta Cryst.* D**65**, 169–175.

Delarue, M. (2008). *Acta Cryst.* D**64**, 40–48.

DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwaï, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature (London)*, **473**, 540–543.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Giacovazzo, C. (2015). *Acta Cryst.* A**71**, 36-45.

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

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.

Glykos, N. M. & Kokkinidis, M. (2004). *J. Appl. Cryst.* **37**, 159–161.

Headd, J. J., Echols, N., Afonine, P. V., Grosse-Kunstleve, R. W., Chen, V. B., Moriarty, N. W., Richardson, D. C., Richardson, J. S. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 381–390.

Jamrog, D. C., Zhang, Y. & Phillips, G. N. (2003). *Acta Cryst.* D**59**, 304–314.

Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 119–124.

Kidera, A. & Gō, N. (1992). *J. Mol. Biol.* **225**, 457–475.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.

Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.

Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.

Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *Proteins*, **40**, 389–408.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443–453.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Rabinovich, D., Rozenberg, H. & Shakked, Z. (1998). *Acta Cryst.* D**54**, 1336–1342.

Schröder, G. F., Brunger, A. T. & Levitt, M. (2007). *Structure*, **15**, 1630–1641.

Schröder, G. F., Levitt, M. & Brünger, A. T. (2010). *Nature (London)*, **464**, 1218–1222.

Schwarzenbacher, R., Godzik, A. & Jaroszewki, L. (2008). *Acta Cryst.* D**64**, 133–140.

Söding, J. (2005). *Bioinformatics*, **21**, 951–960.

Terwilliger, T. C. (2001). *Acta Cryst.* D**57**, 1755–1762.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Acta Cryst.* D**64**, 61–69.

Terwilliger, T. C., Read, R. J., Adams, P. D., Brunger, A. T., Afonine, P. V., Grosse-Kunstleve, R. W. & Hung, L.-W. (2012). *Acta Cryst.* D**68**, 861–870.

Turk, D. (2013). *Acta Cryst.* D**69**, 1342–1357.

Vagin, A. & Teplyakov, A. (2010). *Acta Cryst.* D**66**, 22–25.