Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 21 May 2009
Accepted 3 September 2009

In honour of Herbert Hauptman and Michael Rossmann

Molecular replacement: the probabilistic approach of the program *REMO09* and its applications

Rocco Caliandro, Benedetta Carrozzini, Giovanni Luca Cascarano, Carmelo Giacovazzo,* Annamaria Mazzone and Dritan Siliqi

Institute of Crystallography – CNR, Via G. Amendola, 122/O 70126 Bari, Italy. Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

The method of joint probability distribution functions has been applied to molecular replacement techniques. The rotational search is performed by rotating the reciprocal lattice of the protein with respect to the calculated transform of the model structure; the translation search is performed by fast Fourier transform. Several cases of prior information are studied, both for the rotation and for the translation step: *e.g.* the conditional probability density for the rotation or the translation of a monomer is found both for *ab initio* and when the rotation and/or the translation values of other monomers are given. The new approach has been implemented in the program *REMO09*, which is part of the package for global phasing *IL MILIONE* [Burla, Caliandro, Camalli, Cascarano, De Caro, Giacovazzo, Polidori, Siliqi & Spagna (2007). *J. Appl. Cryst.* **40**, 609–613]. A large set of test structures has been used for checking the efficiency of the new algorithms, which proved to be significantly robust in finding the correct solutions and in discriminating them from noise. An important design concept is the high degree of automatism: *REMO09* is often capable of providing a reliable model of the target structure without any user intervention.

1. Notation

We will use the same notation as in Caliandro *et al.* (2006), hereafter denoted as paper I. Furthermore:

n: number of monomers in the asymmetric unit of the protein and of the model structure.

t: number of atoms in the asymmetric unit of the protein (t/n is the number of atoms per monomer).

p: number of atoms in the asymmetric unit of the model structure (p/n is the number of atoms per monomer).

$\mathbf{r}_j, j = 1, \ldots, t$: atomic positions of the protein structure (symmetry independent).

$\mathbf{r}'_j = \mathbf{r}_j + \Delta\mathbf{r}_j, j = 1, \ldots, p$: atomic positions of the model structure (symmetry independent).

N: number of atoms in the unit cell of the protein structure.

N': number of atoms in the unit cell of the model structure.

$f_j, j = 1, \ldots, t$: atomic scattering factors of the protein, temperature factor included.

$f'_j, j = 1, \ldots, p$: scattering factors of the atoms of the model molecule, temperature factor included.

$\sum_N = \varepsilon \sum_j f_j^2$: the summation is extended to the N atoms of the protein (thermal factors included).

$\sum_{N'} = \varepsilon \sum_j f_j'^2$: the summation is extended to the N' atoms of the model structure (thermal factors included).

$m_1(x) = I_1(x)/I_0(x)\, I_i$: the modified Bessel function of order i.

$D = \langle \cos[2\pi \overline{\mathbf{h}}_{\rm prot} \mathbf{R}_s \Delta\mathbf{r}_j] \rangle_{s,j}$: the average is performed over the $\Delta\mathbf{r}_j$ vectors and on the $\mathbf{R}_s$ matrices. $\mathbf{R}_s$ is the rotational component of the sth symmetry element. D_α, $D_{\alpha,\beta}$ *etc.* denote the values of D calculated for the monomer α, for the monomer pair (α, β) *etc.*, respectively.

NCS: noncrystallographic symmetry.

2. Introduction

Molecular replacement (MR) is one of the most popular techniques for macromolecular phasing. Several programs are today available, based on different theoretical approaches. One way is to directly explore the full six-dimensional space (*e.g.* Chang & Lewis, 1997; Kissinger *et al.*, 1999; Sheriff *et al.*, 1999; Glykos & Kokkinidis, 2000; Jamrog *et al.*, 2003), a choice expensive in terms of computing resources. The six-dimensional search is more frequently broken up into rotation and translation steps (*e.g.* Rossmann & Blow, 1962; Navaza, 1994; Vagin & Teplyakov, 1997; Read, 2001; Yao, 2002; Caliandro *et al.*, 2006; McCoy *et al.*, 2007). The reader will find a comprehensive review of the literature and of the various MR approaches in the January 2008 issue of *Acta Crystallographica Section D*, containing the proceedings of a previous CCP4 study weekend.

In paper I the theoretical approach of the program *REMO* for MR and its first applications were described. In particular:

(*a*) The space group of the model structure was assumed to be the symmorphic variant of the protein space group.

(*b*) A special algebra for the rotation step was outlined, for including the rotational symmetry of the space group during the rotation-function step. In this way all of the orientations of the molecules in the model can be simultaneously superimposed with all the orientations of the actual structure, potentially giving a higher signal (in space groups other than *P*1) compared with standard approaches.

(*c*) The oriented model molecules were located in the unit cell by using correlation functions calculated by fast Fourier transform.

In this paper we will develop the probabilistic approach that the new version of *REMO* (from now on *REMO09*) is based on. As in *REMO*, the MR problem will be subdivided into rotation and translation steps, and for the rotation the special *REMO* algebra will be applied. For this step we will derive joint probability distributions in the absence of or under various prior conditions, *e.g.* the conditional probability density for the rotation of a monomer is found *ab initio* or given the rotation and/or the translation values of other monomers. Since we will suppose, as in paper I, that the rotation search is performed by continuously rotating the indices of the protein structure by means of the rotation matrix $\mathbf{M}_{\rm prot}$, this last one will be a basic parameter of the related distributions.

Joint probability distributions will also be applied to the translation step under various prior conditions. For example, the most probable translation shifts for a given monomer will be obtained when its orientation is known, and also when the orientation and/or the locations of other monomers are known. Since the translation search is performed by continuously moving the model, the translation matrix **N** will be a basic parameter of the related distribution functions. When pseudo-translational symmetry is present, the corresponding information may be actively used in the probabilistic approach.

The derivation of the various conditional distributions (one for each type of prior information) are separately described in the Appendices. For both rotation and translation steps two joint probability distributions were calculated, one for the correct rotation or translation matrix, and one for the incorrect matrices. The comparison of the respective distributions is used to obtain simple and effective (as the test proved) criteria for finding the correct solutions. Any attempt to consider more general distributions, which include as special cases those derived here, lead to non-manageable and complicated mathematical expressions.

It is worthwhile noticing that probabilistic approaches to MR have already been described and implemented in the computing programs *Beast* (Read, 1999) and *Phaser* (McCoy *et al.*, 2007), where maximum-likelihood-based conditional distributions are applied, taking prior information into account. The final formulas derived in this paper do not coincide with those obtained *via* the maximum-likelihood principle and *via* Patterson methods, but are certainly correlated with them. A comparison between the different approaches is beyond the aims of the present paper.

A large number of practical cases will be used to better understand the role of the various parameters included in our probabilistic theory, and to assess the validity of the conclusive formulas. *REMO09*, together with *SAD–MAD*, *SIR–MIR* and *ab initio* techniques, is part of the program *IL MILIONE* (Burla *et al.*, 2007), a general purpose computer package devoted to the global solution of the protein phase problem.

As in paper I, to simplify the mathematical treatment we will assume that both the protein and the model structure contain *n* monomers in the asymmetric unit, referred by noncrystallographic symmetry. The theory, however, may be easily generalized to the case in which monomers of different composition occur.

3. Definitions and algebraic background for the rotation step

Suppose that the asymmetric unit of a protein structure contains *n* monomers referred by NCS, each constituted by *t*/*n* atoms, and that $\boldsymbol{C}_s \equiv (\mathbf{R}_s, \mathbf{T}_s)$, $s = 1, \ldots, m$, are the space group symmetry operators. We will conventionally refer to this structure as the *protein structure*. Its structure factor is

$$F(\mathbf{h}_{\rm prot}) = \sum_{s=1}^{m}\sum_{\mu=1}^{n} a(\overline{\mathbf{h}}_{\rm prot}, \mathbf{T}_s) g_\mu(\overline{\mathbf{h}}_{\rm prot}\mathbf{R}_s), \qquad (1)$$

where

$$a(\overline{\mathbf{h}}_{\rm prot}, \mathbf{T}_s) = \exp(2\pi i\, \overline{\mathbf{h}}_{\rm prot}\mathbf{T}_s),$$
$$g_\mu(\overline{\mathbf{h}}_{\rm prot}\mathbf{R}_s) = \sum_{j=(\mu-1)(t/n)+1}^{\mu t/n} f_j \exp(2\pi i\, \overline{\mathbf{h}}_{\rm prot}\mathbf{R}_s\mathbf{r}_j). \qquad (2)$$

Let us consider another structure (from now on the *model structure*), whose space group is the symmorphic variant of the protein space group: it has only one monomer in the asymmetric unit constituted by *p*/*n* atoms located at $\mathbf{r}'_j, j = 1, \ldots, p/n$. The variables $\mathbf{r}'_j, j = 1, \ldots, p/n$, are correlated with the positions $\mathbf{r}_j$ provided a suitable rotation matrix $\mathbf{M}_{\rm mod}$ and a suitable translation matrix $\mathbf{N}_{\rm mod}$ are applied: *i.e.* $\mathbf{r}_j = \mathbf{M}_{\rm mod}\mathbf{r}'_j + \mathbf{N}_{\rm mod} + \Delta\mathbf{r}'_j$, where $\Delta\mathbf{r}'_j$ are positional shifts sufficiently small to secure isomorphism between the protein and the model molecule, at least at low resolution. Its structure factor is

$$\begin{aligned} F_{\rm mod}(\mathbf{h}_{\rm mod}) &= \sum_{s=1}^{m}\sum_{j=1}^{p/n} f'_j \exp\left[2\pi i\, \overline{\mathbf{h}}_{\rm mod}\mathbf{R}_s\left(\mathbf{M}_{\rm mod}\mathbf{r}'_j + \mathbf{N}_{\rm mod}\right)\right] \\ &= \sum_{s=1}^{m} a(\overline{\mathbf{h}}_{\rm mod}, \mathbf{R}_s\mathbf{N}_{\rm mod})\gamma(\overline{\mathbf{h}}_{\rm mod}\mathbf{R}_s) \end{aligned} \qquad (3)$$

where

$$a(\overline{\mathbf{h}}_{\rm mod}, \mathbf{R}_s\mathbf{N}_{\rm mod}) = \exp(2\pi i\, \overline{\mathbf{h}}_{\rm mod}\mathbf{R}_s\mathbf{N}_{\rm mod}),$$
$$\gamma(\overline{\mathbf{h}}_{\rm mod}\mathbf{R}_s) = \sum_{j=1}^{p/n} f'_j \exp(2\pi i\, \overline{\mathbf{h}}_{\rm mod}\mathbf{R}_s\mathbf{M}_{\rm mod}\mathbf{r}'_j). \qquad (4)$$

Let us now rotate in a continuous way the indices of the protein structure by the matrix $\mathbf{M}_{\rm prot}$. The set of symmetry-equivalent indices $\overline{\mathbf{R}}_\nu\mathbf{h}_{\rm prot}$, $\nu = 1, \ldots, m$, will be transformed into the new indices $\mathbf{M}_{\rm prot}\overline{\mathbf{R}}_\nu\mathbf{h}_{\rm prot}$, $\nu = 1, \ldots, m$, to which the observed moduli $|F(\mathbf{h}_{\rm prot})|^2$ will be constantly associated. If we consider the reflections (for the sake of continuity we use the notation employed in paper I)

$$\mathbf{h}_{\mathrm{mod},s,s} = \mathbf{R}_s \mathbf{M}_{\mathrm{prot}} \overline{\mathbf{R}}_s \mathbf{h}_{\mathrm{prot}}, \quad \text{for } s = 1, \ldots, m, \tag{5}$$

the sum

$$\overline{F} = \sum_{s=1}^{m} |\gamma(\overline{\mathbf{h}}_{\mathrm{mod},s,s} \mathbf{R}_s)|^2 \tag{6}$$

will be correlated with the sum $\sum_{s=1}^{m} |g(\overline{\mathbf{h}}_{\mathrm{prot}} \mathbf{R}_s)|^2$ when $\mathbf{M}_{\mathrm{prot}}$ = $\mathbf{M}_{\mathrm{mod}}$. Indeed,

$$\begin{aligned} \sum_{s=1}^{m} |\gamma(\overline{\mathbf{h}}_{\mathrm{mod},s,s} \mathbf{R}_s \mathbf{M}_{\mathrm{mod}})|^2 &= \sum_{s=1}^{m} |\gamma(\overline{\mathbf{h}}_{\mathrm{prot}} \mathbf{R}_s \overline{\mathbf{M}}_{\mathrm{prot}} \overline{\mathbf{R}}_s \mathbf{R}_s \mathbf{M}_{\mathrm{mod}})|^2 \\ &= \sum_{s=1}^{m} |\gamma(\overline{\mathbf{h}}_{\mathrm{prot}} \mathbf{R}_s \overline{\mathbf{M}}_{\mathrm{prot}} \mathbf{M}_{\mathrm{mod}})|^2 \\ &= \sum_{s=1}^{m} |\gamma(\overline{\mathbf{h}}_{\mathrm{prot}} \mathbf{R}_s)|^2. \end{aligned} \tag{7}$$

Owing to the assumed isomorphism between protein and model structure, when $\mathbf{M}_{\mathrm{prot}}$ = $\mathbf{M}_{\mathrm{mod}}$ one can expect that the correlation between $\overline{F}$ and

$$\begin{aligned} |F(\mathbf{h}_{\mathrm{prot}})|^2 = & \sum_{s=1}^{m} |g(\overline{\mathbf{h}}_{\mathrm{prot}} \mathbf{R}_s)|^2 + \sum_{s_1 \neq s_2 = 1}^{m} \big[a(\overline{\mathbf{h}}_{\mathrm{prot}}, \mathbf{T}_{s_1} - \mathbf{T}_{s_2}) \\ & \times g(\overline{\mathbf{h}}_{\mathrm{prot}} \mathbf{R}_{s_1}) g(-\overline{\mathbf{h}}_{\mathrm{prot}} \mathbf{R}_{s_2})\big] \end{aligned} \tag{8}$$

is larger than for a generic rotation $\mathbf{M}_{\mathrm{prot}}$. Indeed, in accordance with the protein space-group symmetry, the rotation $\mathbf{M}_{\mathrm{prot}}$ = $\mathbf{M}_{\mathrm{mod}}$ has brought a molecule of the protein structure into the same orientation as a model molecule: simultaneously, any molecule of the model will be iso-oriented with another molecule of the model.

Owing to the central role of $\overline{F}$, it is important to establish its statistical behaviour: this is done in §4.

4. The distribution $P(\overline{E})$

Suppose that the asymmetric unit of the model structure is constituted by a single molecule of p atoms. In accordance with (6), $\overline{F}$ is real and non-negative: we will calculate the probability distribution $P(\overline{F})$ under the following assumptions:

(a) the coordinates of the vectors $\mathbf{r}'_j$, j = 1,..., p, are the primitive random variables of our approach, uniformly distributed in the unit cell.

(b) γ_μ is statistically independent of γ_ω for $\mu \neq \omega$.

We apply a property of the γ-functions (Srinivasan & Subramanian, 1964; see also Shmueli & Wilson, 1993): if $x_1, x_2, \ldots, x_m$ are independent γ-distributed variables with parameters $n_1, n_2, \ldots, n_m$, their sum is a γ-distributed variable with n = $n_1 + n_2 + \ldots + n_m$. Since the variables $(\gamma'_s)^2$ = $\gamma_s^2 / (\sum_{j=1}^{p} f_j'^2)$ are γ-distributed with n_s = 1 for any s, then $\overline{E}'$ = $\sum_{s=1}^{m} \gamma_s'^2$ is γ-distributed with parameter m:

$$P(\overline{E}') = [\Gamma(m)]^{-1} \, \overline{E}'^{m-1} \exp(-\overline{E}').$$

If we introduce the change of variable

$$\overline{E} = \sum_{s=1}^{m} \gamma_s^2 / \textstyle\sum_{N'} = \overline{E}'/m,$$

we have

$$P(\overline{E}) = [\Gamma(m)]^{-1} \, m^m \, \overline{E}^{m-1} \exp(-m\overline{E}), \tag{9}$$

which is shown in Fig. 1 for m = 1, 2, 4, 8. Different values of m imply different distributions: simple calculations show that

$$\langle \overline{E} \rangle = 1 \tag{10}$$

for any m, but

$$\langle \overline{E}^2 \rangle = (m + 1)/m. \tag{11}$$

The value of $\langle \overline{E}^2 \rangle$ is therefore space-group dependent. The averages (10) and (11) will be used to derive distributions useful for the probabilistic treatment of the rotation step.

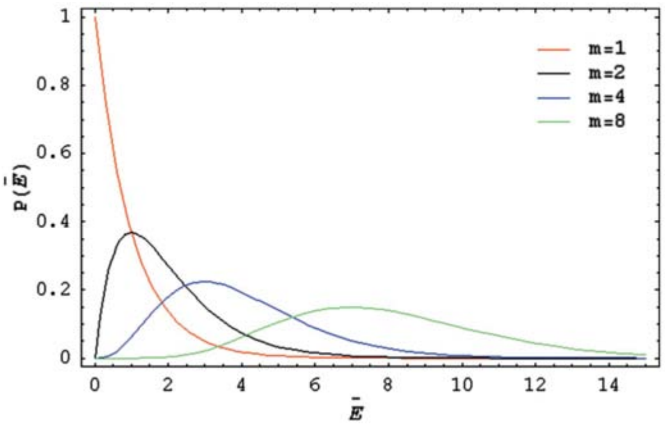

Figure 1
The probability distribution $P(\overline{E})$ = $[\Gamma(m)]^{-1} m^m \overline{E}^{m-1}$ $\exp(-m\overline{E})$ for m = 1, 2, 4, 8.

5. The rotation step: the distribution $P(E, \overline{E})$

Suppose that the asymmetric unit of the protein structure is constituted by t atoms organized in n monomers, and that the asymmetric unit of the model structure contains only one monomer. In our probabilistic approach the variable $F \equiv F(\mathbf{h}_{\mathrm{prot}})$ is defined by equations (1)–(2), and E = $A + iB = R\exp(i\phi)$ is the corresponding normalized structure factor. We will first study the distribution $P(E, \overline{E})$ under the following assumptions:

(a) $\mathbf{M}_{\mathrm{prot}}$ = $\mathbf{M}_{\mathrm{mod}}$: then

$$\overline{F} = \sum_{s=1}^{m} \gamma_s^2, \qquad \gamma_s^2 = \left| \sum_{j=1}^{p/n} f_j' \exp(2\pi i \, \overline{\mathbf{h}}_{\mathrm{prot}} \mathbf{R}_s \mathbf{r}'_j) \right|^2.$$

The distribution $P(E, \overline{E})$ when $\mathbf{M}_{\mathrm{prot}} \neq \mathbf{M}_{\mathrm{mod}}$ will be easily obtained from the previous case.

(b) The coordinates of the vectors $\mathbf{r}_j, j = 1, \ldots, t$, are the primitive random variables of our approach, uniformly distributed in the unit cell.

(c) The variables $\mathbf{r}'_j, j = 1, \ldots, p/n$, are riding variables. They are uncorrelated with the corresponding $\mathbf{r}_j$ owing to the fact that the position of the model molecule is unknown, but the interatomic distances $\mathbf{r}'_i - \mathbf{r}'_j$ are correlated with the vectors $\mathbf{r}_i - \mathbf{r}_j$ through the local positional errors $\Delta\mathbf{r}'_j$ for $j = 1, \ldots, p/n$.

We obtain (see Appendix *A*)

$$P(R,\overline{E}) = (2/\pi)^{1/2} m^{1/2} R \exp\left[-R^2 - \tfrac{1}{2}m(\overline{E}-1)^2\right] \times \left[1 + 2mk_{201}(\overline{E}-1)(R^2-1)\right]. \qquad (12a)$$

where

$$2mk_{201} = \sigma_A^2, \qquad \sigma_A^2 = D^2 \frac{\sum_{N'/n}}{\sum_N} \qquad \text{if } p \le t,$$

$$\sigma_A^2 = \frac{1}{n^2} D^2 \frac{\sum_N}{\sum_{N/n'}} \qquad \text{if } p > t.$$

The term D takes into account the mismatch between model and protein molecules. The distribution (12*a*) is space-group dependent (it depends on the m value): when $\mathbf{M}_{\rm prot} \neq \mathbf{M}_{\rm mod}$ it reduces to

$$P(R,\overline{E}) = (2/\pi)^{1/2} m^{1/2} R \exp\left[-R^2 - \tfrac{1}{2}m(\overline{E}-1)^2\right]. \qquad (12b)$$

The comparison of equation (12*a*) with (12*b*) suggests that the factor $\sigma_A^2(\overline{E}-1)(R^2-1)$ makes the difference between the two distributions. Accordingly, when the joint probability distribution function of several pairs $(E,\overline{E})$ is taken into account, one can use

$$\textstyle\sum_i \sigma_A^2(\overline{E}_i-1)(R_i^2-1) = \max$$

as a criterion for identifying the rotations for which $\mathbf{M}_{\rm prot} \simeq \mathbf{M}_{\rm mod}$.

The above probabilistic results and the following observations allow us to design a simpler figure of merit for identifying the correct rotations:

(*a*) σ_A^2 constitutes a possible statistical criterion to evaluate *a priori* the difficulty of a specific rotation problem. Indeed $2k_{201}$ coincides with the average $\langle(\overline{E}-1)(R^2-1)\rangle$, expected when the correct rotation is used,

$$\langle(\overline{E}-1)(R^2-1)\rangle \simeq \langle\overline{E}R^2\rangle - 1 = 2m_{201} - 1 = 2k_{201}.$$

Difficult rotation searches are expected to occur when the number of monomers in the asymmetric unit is large, and/or when the ratio between model and protein molecule size largely differs from unity (*i.e.* when N' is quite different from N), and/or when there is a strong mismatch between model and protein structure similarity (say, $D \ll 1$). Most difficult cases are characterized by values of σ_A^2 close to zero.

(*b*) In principle, σ_A^2 and therefore D^2 may be estimated *via* a statistical analysis of the pairs $(\overline{E}, R^2)$ (see Luzzati, 1952; Srinivasan & Ramachandran, 1965; Read, 1986). We used the statistical approach described by Caliandro *et al.* (2005), which takes into account the measurement error $\langle|\mu|^2\rangle$: accordingly,

$$\sigma_A^2 = \langle\overline{E}R^2\rangle - e$$

where $e = (1 + \langle|\mu|^2\rangle / \sum_N)$. Such an approach, however, was not fruitful: in fact σ_A^2 does not show a well defined trend *versus* the resolution, owing to the low correlation between $\overline{E}$ and R^2 (that is not unexpected: the first parameter depends on the intramolecular vectors only, the second on both intra- and intermolecular vectors). Such a result discouraged the use of the resolution-dependent σ_A parameter: in practice, in our calculations any constant value may be used. It may be worthwhile noting that *Phaser* uses the prior estimate from the sequence identity as a supplementary source of prior information.

(*c*) Owing to the similarity between the model and the target structures, the ratio $\sum_{N'} / \sum_N$ may be approximated by the same constant for different resolution shells. To spare computing time, it may be excluded from the criterion.

(*d*) The multiplicity of each reflection may be included in the criterion, to make it more robust.

In accordance with the above observations, computing time may be spared without remarkable loss of efficiency if the criterion

$$\mathrm{FOM}_R = \textstyle\sum_i M_{ui}(\overline{E}_i - 1)(R_i^2 - 1) = \max \qquad (13)$$

is used, where M_u is the reflection multiplicity. The probabilistic nature of our approach suggested the elimination, from the right-hand side of (13), of the subset of reflections for which both R and $\overline{E}$ are too close to unity (*i.e.* between 0.7 and 1.35): in this way only the reflections providing a large contribution are included in the summation. The orientations corresponding to the highest values of FOM_R may be selected according to the normalized variable

$$\rho_R = \frac{\mathrm{FOM}_{R\,\max} - \mathrm{FOM}_R}{\mathrm{FOM}_{R\,\max} - \mathrm{FOM}_{R\,\min}}. \qquad (14)$$

A selection threshold may be applied taking into account the expected difficulty of the rotation case: in particular, when $n >$ 1 and the sequence identity is low the default choice requires that at least 20 possible solutions are in the selected set. The selected rotations are refined by performing a finer rotational search within a neighbourhood (in the orientation space) of the given rotation (see paper I for further details). Refined solutions are clustered. The representative orientations remaining after the clustering analysis are further on selected through the criterion (14): the new locally optimized values of FOM_R are used.

6. Orienting a monomer when one or more other monomers have already been oriented

Let us suppose that the asymmetric unit of the model structure is constituted of n monomers related by NCS, and that we have already found the orientation of the first monomer, the monomer α, constituted of p/n atoms. Let

$$\overline{F}_\alpha = \sum_{s=1}^{m} \left| \sum_{j=1}^{p/n} f_j' \exp(2\pi i\, \overline{\mathbf{h}}_{\rm prot} \mathbf{R}_s \mathbf{r}_j') \right|^2$$

be the value of $\overline{F}_\alpha$ when α has been correctly oriented. We want to search for the orientation of a second monomer, the monomer β. We will study the distribution $P(E,\overline{E})$ where the variable $F \equiv F(\mathbf{h}_{\rm prot})$ is defined by equation (1) under the supplementary condition that $\overline{F}_\alpha$ is *a priori* known, and

$$\overline{F} = \overline{F}_{\alpha+\beta} = \overline{F}_\alpha + \overline{F}_\beta.$$

The correct orientation will be found when $\mathbf{M}_{\rm prot} = \mathbf{M}_{\beta\,\rm mod}$.

The joint distribution $P(E, \overline{E})$ will be calculated under the following conditions:

(*a*) $\overline{F}_\alpha$, as well as the interatomic vectors $\mathbf{r}'_i - \mathbf{r}'_j$, $i, j = 1, \ldots, p/n$, are fixed parameters of the distribution.

(*b*) The coordinates of the vectors $\mathbf{r}_j$, $j = 1, \ldots, p/n$, are riding variables: they are uncorrelated with the corresponding $\mathbf{r}'_j$ owing to the fact that the position of the model molecule is unknown, but the interatomic distances $\mathbf{r}_i - \mathbf{r}_j$ are correlated with the vectors $\mathbf{r}'_i - \mathbf{r}'_j$ through the local positional errors $\Delta\mathbf{r}_j$.

(*c*) The coordinates of the vectors $\mathbf{r}_j$, $j = p/n + 1, \ldots, t$ are primitive random variables.

(*d*) The variables $\mathbf{r}'_j$, $j = p/n + 1, \ldots, 2p/n$, are riding variables: they are correlated with the corresponding $\mathbf{r}_j$ through the local positional errors $\Delta\mathbf{r}_j$.

Let $E = F/\sum_N^{1/2}$, and let $\overline{E} = \overline{F}/\sum_{N'/n}$ be a pseudo-normalized structure factor. From Appendix *B*, equation (46), we obtain

$$P(R, \overline{E}) \simeq R \exp\left[-\frac{R^2}{2k_{200}} - \tfrac{1}{2}m(\overline{E} - \overline{E}_\alpha - 1)^2\right] \times \left[1 + m\frac{k_{201}}{2k_{200}}(\overline{E} - \overline{E}_\alpha - 1)\left(\frac{R^2}{2k_{200}} - 1\right)\right] \qquad (15)$$

where

$$k_{200} = \frac{1}{2}\left[\frac{1}{n}D_\alpha^2\overline{E}_\alpha\frac{\sum_{N'}}{\sum_N} + \frac{(n-1)}{n}\right],$$

$$k_{201} = \frac{1}{2m}D_\beta^2\frac{\sum_{N'/n}}{\sum_N} \qquad \text{if } p \le t \qquad \text{and}$$

$$k_{201} = \frac{1}{2mn^2}D_\beta^2\frac{\sum_N}{\sum_{N'/n}} \qquad \text{if } p > t.$$

Since $\overline{E} = \overline{E}_\alpha + \overline{E}_\beta$, equation (15) may be simplified as follows,

$$P(R, \overline{E}_\beta) \simeq R \exp\left[-\frac{R^2}{2k_{200}} - \tfrac{1}{2}m(\overline{E}_\beta - 1)^2\right] \times \left[1 + m\frac{k_{201}}{2k_{200}}(\overline{E}_\beta - 1)\left(\frac{R^2}{2k_{200}} - 1\right)\right]. \qquad (16)$$

When $\mathbf{M}_{\rm prot} \neq \mathbf{M}_{\rm mod}$, equation (16) reduces to

$$P(R, \overline{E}_\beta) \simeq (2/\pi)^{1/2}m^{1/2}R \exp\left[-\frac{R^2}{2k_{200}} - \tfrac{1}{2}m(\overline{E}_\beta - 1)^2\right].$$

In accordance with §5, when the joint probability distribution function of several pairs $(E, \overline{E})$ is taken into account, one can use

$$\textstyle\sum_i (k_{200})^{-2} m k_{201}(\overline{E}_{\beta i} - 1)(R_i^2 - 2k_{200}) = \max \qquad (17)$$

as a criterion for identifying the correct orientation of the second monomer.

The results obtained above may be easily generalized to the case in which two monomers have already been oriented, and we want to find the orientation of a third one. Then we should again study the distribution $P(E, \overline{E})$, where

$$\overline{F} = \overline{F}_{\alpha+\beta+\gamma} = \overline{F}_\alpha + \overline{F}_\beta + \overline{F}_\gamma,$$

$$\overline{F} = F_{\rm mod}(\mathbf{h}_{\rm mod}) = F_\gamma = \sum_{s=1}^{m}\left|\sum_{j=2p/n+1}^{3p/n} f_j' \exp(2\pi i\,\overline{\mathbf{h}}_{\rm prot}\mathbf{R}_s\mathbf{r}'_j)\right|^2.$$

We will assume that the orientation of the first monomer has already been found when $\mathbf{M}_{\rm prot}$ = $\mathbf{M}_{\alpha\,{\rm mod}}$, and that of the second when $\mathbf{M}_{\rm prot}$ = $\mathbf{M}_{\beta\,{\rm mod}}$. That makes available the prior information

$$\overline{F}_\alpha = \sum_{s=1}^{m}\left|\sum_{j=1}^{p/n} f_j' \exp(2\pi i\,\overline{\mathbf{h}}_{\rm prot}\mathbf{R}_s\mathbf{r}'_j)\right|^2 \qquad \text{and}$$

$$\overline{F}_\beta = \sum_{s=1}^{m}\left|\sum_{j=p/n+1}^{2p/n} f_j' \exp(2\pi i\,\overline{\mathbf{h}}_{\rm prot}\mathbf{R}_s\mathbf{r}'_j)\right|^2.$$

Then the distribution

$$P(R, \overline{E}_\gamma) \simeq (2/\pi)^{1/2}m^{1/2}R \exp\left[-\frac{R^2}{2k_{200}} - \tfrac{1}{2}m(\overline{E}_\gamma - 1)^2\right] \times \left[1 + m\frac{k_{201}}{2k_{200}}(\overline{E}_\gamma - 1)\left(\frac{R^2}{2k_{200}} - 1\right)\right] \qquad (18)$$

is obtained, where

$$k_{200} = \frac{1}{2}\left[(D_\alpha^2\overline{E}_\alpha + D_\beta^2\overline{E}_\beta)\frac{\sum_{N'}}{\sum_N} + \frac{(n-2)}{n}\right],$$

$$k_{201} = \frac{1}{2m}D_\gamma^2\frac{\sum_{N'/n}}{\sum_N} \qquad \text{if } p \le t \qquad \text{and}$$

$$k_{201} = \frac{1}{2mn^2}D_\gamma^2\frac{\sum_N}{\sum_{N'/n}} \qquad \text{if } p > t.$$

The relation

$$\textstyle\sum_i (k_{200})^{-2} m k_{201}(\overline{E}_{\gamma i} - 1)(R_i^2 - 2k_{200}) = \max \qquad (19)$$

may then be used as a criterion for identifying the orientation of the third monomer. The extension of the procedure for the search of a next monomer is trivial.

The above probabilistic results and the following observations allow us to design a simpler figure of merit for identifying the correct rotations (we take into consideration the case in which only the monomer α is correctly oriented: the results are easily generalized):

(*a*) $2mk_{201} = (\sigma_A^2)_\beta$, $(\sigma_A^2)_\beta = D_\beta^2\sum_{N'/n}/\sum_N$ if $p \le t$, $(\sigma_A^2)_\beta = (D_\beta^2/n^2)\sum_N/\sum_{N'/n}$ if $p > t$.

(*b*) In agreement with §5, $(\sigma_A^2)_\beta(k_{200})^{-2}$ is a possible statistical parameter estimating the difficulty of orienting the monomer β given the orientation of the monomer α.

(*c*) Both D_α and D_β are settled to 1; then $(\sigma_A^2)_\beta$ may be considered constant and omitted from the calculations.

(*d*) The prior information about the orientation of the monomer is contained in the parameter k_{200}, which, in turn, depends on the normalized structure factor $\overline{E}_\alpha$ and on the parameter $(\sigma_A^2)_\alpha$.

(*e*) $2k_{200}$ is the expected value of R^2 when the prior information is taken into account. Accordingly, in (17), R^2 is compared with $2k_{200}$ rather than with unity, as it occurs in (13).

(*f*) The value $(k_{200})^{-2}$ modulates the contribution arising from each product $(\overline{E}_\beta - 1)(R^2 - 2k_{200})$. Large values of k_{200} and therefore of $\overline{E}_\alpha$ deplete the value of the contribution.

(*g*) The subset of reflections for which both $R/2k_{200}$ and $\overline{E}_\beta$ are between 0.7 and 1.35 are excluded from the calculations.

In accordance with the above observations, computing time may be spared without remarkable loss of efficiency if the criterion

$$\textstyle\sum_h (k_{200})^{-2} M_u(\overline{E}_\beta - 1)(R^2 - 2k_{200}) = \max \qquad (20)$$

is used. A two-step procedure may be devised for recognizing the correct orientation of the second monomer given that of the first one:

(i) The rotations selected by equation (20) are combined in pairs, and, for each pair, the left-hand side of equation (20) is calculated.

(ii) The pairs with the highest score are submitted, in score order, to the translation step.

In case we need to orient a third monomer, equation (19) may be applied, with approximations similar to those described in points (*a*)–(*g*).

7. Translate a well oriented monomer

Let us suppose that the asymmetric unit of the protein unit cell contains n monomers and that the monomer α of the model structure has been correctly oriented. We rewrite the one-monomer model structure factor $\overline{F}$ as

$$\overline{F} = \sum_{s=1}^{m} a_{\alpha,s}, \gamma_{\alpha,s} \qquad (21)$$

where

$$a_{\alpha,s} = \exp[2\pi i\,\overline{\mathbf{h}}(\mathbf{R}_s\mathbf{N}_\alpha + \mathbf{T}_s)], \quad \gamma_{\alpha,s} = \sum_{j=1}^{p/n} f_j{}' \exp(2\pi i\,\overline{\mathbf{h}}\mathbf{R}_s\mathbf{r}'_j).$$

$\mathbf{N}_\alpha$ is the translation vector we are looking for. The protein structure factor may be written as

$$F = \sum_{s=1}^{m}\sum_{\mu=1}^{n} a_s g_{\mu,s}, \qquad (22)$$

where

$$a_s = \exp(2\pi i\,\overline{\mathbf{h}}\mathbf{T}_s), \quad g_{\mu,s} = \sum_{j=(\mu-1)(t/n)+1}^{\mu t/n} f_j \exp[2\pi i\,\overline{\mathbf{h}}\mathbf{R}_s(\mathbf{r}'_j + \Delta\mathbf{r}_j)].$$

The joint probability distribution function $P(E, \overline{E})$ should be calculated under the following conditions:

(*a*) The coordinates of the vectors $\mathbf{r}'_j$, $j = 1, \ldots, p/n$ are fixed parameters of our probabilistic approach, while $\mathbf{N}_\alpha$ is a primitive random vector.

(*b*) The variables $\mathbf{r}_j = \mathbf{r}'_j + \Delta\mathbf{r}_j$, $j = 1, \ldots, p/n$, are riding variables. They are uncorrelated with the corresponding $\mathbf{r}_j$ owing to the fact that the position of the model molecule is unknown, while the interatomic distances $\mathbf{r}'_i - \mathbf{r}'_j$ are correlated with the vectors $\mathbf{r}_i - \mathbf{r}_j$ through the local positional errors $\Delta\mathbf{r}_j$.

(*c*) The variables $\mathbf{r}_j$, for $j = p/n + 1, \ldots, t$, are primitive random variables.

When $\mathbf{N}_\alpha = 0$ the probability distribution refers two structure factors relative to two isomorphous structures (Srinivasan & Ramachandran, 1965; Read, 1986, 1990; Caliandro *et al.*, 2005):

$$P(R, \overline{R}, \phi, \overline{\phi}) = R\overline{R}\pi^{-2}(e - \sigma_A^2)^{-1} \exp\left\{ -\frac{1}{(e - \sigma_A^2)}\Big[R^2 + e\overline{R}^2 - 2\sigma_A R\overline{R}\cos(\phi - \overline{\phi})\Big]\right\}. \qquad (23)$$

The degree of isomorphism depends on the parameter σ_A: perfect isomorphism occurs when $\sigma_A = 1$, the value $\sigma_A = 0$ characterizes two uncorrelated structures. From (23) the following conditional distribution arises,

$$P(\phi - \overline{\phi} | R, \overline{R}, \overline{\phi}) = [2\pi I_0(X)]^{-1} \exp[X\cos(\phi - \overline{\phi})]. \qquad (24)$$

If the reflections are assumed to be independent, the total probability relative to a subset of structure factors is the product of the values (24),

$$P\big[(\phi_i - \overline{\phi}_i)|(R_i, \overline{R}_i, \overline{\phi}_i)\big] = \Big\{\prod_i [2\pi I_0(X_i)]^{-1}\Big\} \times \exp\big[\textstyle\sum_i X_i\cos(\phi_i - \overline{\phi}_i)\big].$$

In particular, when the monomer is correctly located, the following relation is expected,

$$\textstyle\sum_i X_i\cos(\phi_i - \overline{\phi}_i) \simeq \sum_i X_i m_1(X_i).$$

When $\mathbf{N}_\alpha \neq 0$ the model phases are expected to be uncorrelated with the phases of the target structure. Then

$$P(\phi - \overline{\phi}|R, \overline{R}, \overline{\phi}) = (2\pi)^{-1},$$

and, as a consequence (in this case σ_A and the corresponding X parameter are expected to vanish),

$$\textstyle\sum_i X_i\cos(\phi_i - \overline{\phi}_i) \simeq 0.$$

The comparison of the expected cosine invariants at the correct and at an incorrect position suggests the criterion

$$\textstyle\sum_i X_i m_1(X_i) = \max \qquad (25)$$

to recognize the correct location.

The above criterion has an algebraic counterpart in a property of the cross-correlation function between target [say $\rho(\mathbf{r})$] and model [say $\overline{\rho}(\mathbf{r})$] structure. Let us denote by $C(\mathbf{u})$ such cross-correlation,

$$C(\mathbf{u}) = \rho(\mathbf{r}) \otimes \overline{\rho}(\mathbf{r}) = \int_S \rho(\mathbf{r})\overline{\rho}(\mathbf{r} + \mathbf{u})\,\mathrm{d}\mathbf{r} = (1/V)\textstyle\sum_{\mathbf{h}} |F_{\mathbf{h}}\overline{F}_{\mathbf{h}}| \exp i(\phi_{\mathbf{h}} - \boldsymbol{\phi}_{\mathbf{h}})\exp(-2\pi i\,\mathbf{hu}).$$

The crystallographic properties of this function and its usefulness for the phase problem have been recently described by Carrozzini *et al.* (2009): in particular the maxima of $C(\mathbf{u})$ lie at the vectors between model and target atoms. In the origin, $C(\mathbf{u})$ takes the value

$$C(\mathbf{0}) = \textstyle\sum_{\mathbf{h}} |F_{\mathbf{h}}\overline{F}_{\mathbf{h}}| \exp i(\phi_{\mathbf{h}} - \boldsymbol{\phi}_{\mathbf{h}}).$$

This peak corresponds to almost vanishing distances between model and target atoms. $C(\mathbf{0})$ is not a peak if model and target structures are uncorrelated: its intensity increases when the scattering power of the model atoms closely overlapping the corresponding target atoms increases. In the molecular replacement translation step it is expected that the maximum overlapping is obtained when the model structure is shifted into the correct position.

Since the phases $\phi_{\mathbf{h}}$ are unknown, $C(\mathbf{u})$ cannot be calculated for each trial translation, but it may be approximated by

$$C'(\mathbf{u}) = (1/V)\textstyle\sum_{\mathbf{h}} |F_{\mathbf{h}}\overline{F}_{\mathbf{h}}| m_{1\mathbf{h}} \exp(-2\pi i\,\mathbf{hu})$$

which, at the origin, takes the value

$$C'(\mathbf{0}) = \textstyle\sum_{\mathbf{h}} |F_{\mathbf{h}}\overline{F}_{\mathbf{h}}| m_{1\mathbf{h}}.$$

In the MR translation step the maximum at the origin of the cross-correlation map may be used to discriminate the correct from the trial translations *via* the condition $C'(\mathbf{0})$ = max, characterizing the maximum overlap between model and target atoms. If the cross-correlation function is calculated in terms of normalized structure factors then we obtain

$$\textstyle\sum_{\mathbf{h}} |E_{\mathbf{h}}\overline{E}_{\mathbf{h}}| m_{1\mathbf{h}} = \max.$$

This relation, derived by grafting a probabilistic relation into algebraic techniques, cannot coincide with relationship (25), but provides support to it by creating a link with a physical property of the cross-correlation map.

Equation (25) may be simplified to

$$T = \textstyle\sum_i M_{ui} X_i m_1(X_i) = \max. \tag{26}$$

The probabilistic nature of the T criterion suggests considering only reflections for which $X > 1$ [they give the largest contribution to the sum and make closer the relation between expected and real values of $\cos(\phi_i - \phi_{pi})$]. A relevant point to stress is the fact that using $\sum_i X_i m_1(X_i)$ is not equivalent to using $\langle X m_1(X)\rangle$. Indeed, for the correct translation it is expected that the number of reflections for which R and $\mathbf{R}$ are both large or small is bigger than for a trial translation. Dividing $\sum_i X_i m_1(X_i)$ by the number of terms in the summation would deplete the score of the correct translation.

As for *REMO*, a first peak selection is made on the basis of the peak height: to the selected peaks the criterion (26) is applied. In practice the feasible translation vectors are ranked by the criterion

$$\rho_T = \frac{T_{\max} - T}{T_{\max} - R_{\min}}.$$

8. Orient a monomer when one or more other monomers have already been oriented and located

Let us suppose that the asymmetric units of the model and of the protein structure are constituted by n monomers related by NCS and that the orientation and the position of the first monomer (the monomer α) has been found. We are looking for the orientation of a second monomer (the monomer β). In this situation we can assume that

$$F_\alpha = \sum_{s=1}^{m}\sum_{j=1}^{p/n} f_j' \exp[2\pi i\,\overline{\mathbf{h}}_{\rm prot}(\mathbf{R}_s\mathbf{r}_j' + \mathbf{T}_s)]$$

is *a priori* known. The protein structure factor can then be represented as

$$F = F_\alpha + F_{c\alpha},$$

where $F_{c\alpha}$ is the structure factor corresponding to the rest of the protein structure.

We want to search for the orientation of the monomer β, which will be found when $\mathbf{M}_{\rm prot}$ = $\mathbf{M}_{\beta\,\rm mod}$. In this situation the variable

$$\begin{aligned}\overline{F} &= F_{\rm mod}(\mathbf{h}_{\rm mod}) = F_\beta \\ &= \sum_{s=1}^{m}\left|\sum_{j=p/n+1}^{2p/n} f_j' \exp(2\pi i\,\overline{\mathbf{h}}_{\rm prot}\mathbf{R}_s\mathbf{r}_j')\right|^2\end{aligned} \tag{27}$$

may be associated with the monomer β, and the joint probability distribution $P(E, \overline{E})$ may be calculated (the case $\mathbf{M}_{\rm prot} \neq \mathbf{M}_{\beta\,\rm mod}$ may be easily derived from the previous distribution). We will make the following assumptions:

(*a*) The coordinates of the vectors $\mathbf{r}_j'$, $j = 1, \ldots, p/n$, are fixed parameters. The coordinates of the vectors $\mathbf{r}_j$, $j = 1, \ldots, t/n$, are riding variables: they are correlated with the corresponding $\mathbf{r}_j'$ through the local positional errors $\Delta\mathbf{r}_j$.

(*b*) The coordinates of the vectors $\mathbf{r}_j'$, $j = p/n + 1, \ldots, 2p/n$, are primitive random variables.

(*c*) The coordinates of the vectors $\mathbf{r}_j$, $j = t/n + 1, \ldots, 2t/n$, are constrained variables. They are uncorrelated with the corresponding $\mathbf{r}_j'$ owing to the fact that the position of the model molecule is unknown, but the interatomic distances $\mathbf{r}_i - \mathbf{r}_j$ are correlated with the vectors $\mathbf{r}_i' - \mathbf{r}_j'$ through the local positional errors $\Delta\mathbf{r}_j$.

(*d*) The coordinates of the vectors $\mathbf{r}_j$, $j = 2t/n + 1, \ldots, t$, are primitive random variables.

From Appendix *C* we obtain

$$\begin{aligned}P(R', \overline{E}_\beta) \simeq{}& (2/\pi)^{1/2} m^{1/2} R' \\ &\times \exp\{-[R'^2 + D_\alpha^2 R_\alpha'^2 + m(\overline{E}_\beta - 1)]\} \\ &\times I_0(2D_\alpha R' R_\alpha') \\ &\times \{1 + 2k_{201} m(\overline{E}_\beta - 1)[R'^2 + D_\alpha^2 R'_{\alpha^2} \\ &- 2D_\alpha R' R_\alpha' m_1(2D_\alpha R' R_\alpha') - 1]\},\end{aligned} \tag{28}$$

where $\overline{E}_\beta$, R' and R_α' are pseudo-normalized (with respect to the unlocated electron density) structure factors, corresponding to $\overline{F}$, F and F_α, respectively,

$$k_{201} = \frac{1}{2m} D_\beta^2 \frac{\sum_{N'/n}}{\sum_N} \quad \text{if } p \le t, \qquad \text{and}$$

$$k_{201} = \frac{1}{2mn^2}\; D_\beta^2\; \frac{\sum_N}{\sum_{N'/n}} \qquad \text{if } p > t.$$

D_β is the value of D calculated for the monomer β.

When $\mathbf{M}_{\rm prot} \neq \mathbf{M}_{\beta\rm mod}$ the term k_{201} is expected to vanish. Accordingly, the criterion

$$\textstyle\sum_i 2mk_{201}(\overline{E}_{\beta i} - 1)[R_i'^2 + D_\alpha^2 R_{\alpha i}'^2 - X_i' m_1(X_i') - 1] = \max, \tag{29}$$

where X_i' = $2D_\alpha R_i' R_{\alpha i}'$ may be used for discriminating the correct rotation. The left-hand side of (29) may be geometrically interpreted in accordance with Fig. 2, where we plot

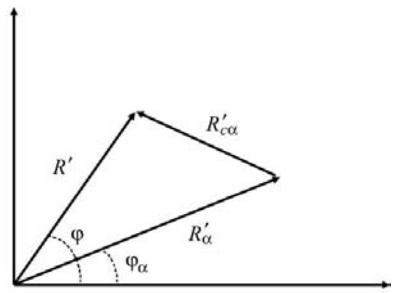

Figure 2
R' and R'_α are pseudo-normalized (with respect to the unlocated electron density) structure factors, $R'_{c\alpha}$ is the pseudo-normalized difference structure factor modulus.

R', R'_α and the pseudo-normalized difference structure factor modulus $R'_{c\alpha}$. In accordance with the Carnot theorem,

$$|R'_{c\alpha}|^2 = |R'|^2 + |R'_\alpha|^2 - 2|R'R'_\alpha|\cos(\varphi - \varphi_\alpha).$$

Given the information on φ_α, and in the absence of information on φ, the expected value of $|R'_{c\alpha}|^2$ is

$$\langle |R'_{c\alpha}|^2\rangle = |R'|^2 + |R'_\alpha|^2 - 2|R'R'_\alpha| m_1(2R'R'_\alpha),$$

where $m_1(x) = I_1(x)/I_o(x)$ is the expected value of $\cos(\varphi - \varphi_\alpha)$ and I_i is the modified Bessel function of order i. Accordingly, (29) may be approximately rewritten as

$$\textstyle\sum 2mk_{201}(\overline{E}_\beta - 1)(\langle R'^2_{c\alpha}\rangle - 1) = \max.$$

The geometrical interpretation suggests that the criterion (29) reduces the complexity of the search by taking care of the prior information: since the value of $R^2_{c\alpha}$ is unknown, the criterion uses its expected value. To spare computing time, the left-hand side of (29) may be expressed in terms of normalized structure factors $R_\alpha = F_\alpha / \sum^{1/2}_{N'/n}$ and $R = F / \sum^{1/2}_N$ (already calculated in the preceding steps of the *REMO09* procedure) as follows. Since

$$R'_\alpha = Q_a R_\alpha, \qquad R' = QR,$$

where $Q_a = \sum^{1/2}_{N'/n} / [\sum_N - \sum_{N'/n}]^{1/2}$, $Q = \sum^{1/2}_N / [\sum_N - \sum_{N'/n}]^{1/2}$, we can rewrite (29) as

$$\textstyle\sum_i 2mk_{201}(\overline{E}_{\beta i} - 1)\big\{K[R^2_i + (\sigma^2_A)_{\alpha i} R^2_{\alpha i} - X_i m_1(KX_i)] - 1\big\} = \max,$$

where $X_i = 2\sigma_{\alpha i} R_i R_{\alpha i}$ and $K = \sum_N / (\sum_N - \sum_{N'/n})$.

In accordance with the previous sections,

(*a*) We do not calculate k_{201} (it is assumed to be roughly constant against resolution).

(*b*) We use the value $D = 1$.

(*c*) Only the reflections with $KX > 1$ are involved in the calculations [for these the probability that $m_1(KX_i)$ is the expected value of $\cos(\varphi - \varphi_\alpha)$ is higher].

(*d*) The reflection multiplicity is introduced.

The actual criterion to be tested is therefore

$$\textstyle\sum_i M_{ui}(\overline{E}_{\beta i} - 1)(\langle R^2_{c\alpha i}\rangle - 1) = \max, \tag{30}$$

where

$$\langle R^2_{c\alpha i}\rangle = K[R^2_i + (\sigma^2_A)_{\alpha i} R^2_{\alpha i} - 2X_i m_1(2X_i)].$$

The procedure for recognizing the correct orientation of the second monomer given the correct orientation and translation of the first may be described by two steps:

(i) The orientation of the first monomer is combined in pairs with the rotations selected (in the absence of any prior information) by the procedure described in §4 and, for each pair, the left-hand side of equation (30) is calculated.

(ii) The pairs with the largest score are submitted, in score order, to the translation step as described in §7 (or to a supplementary procedure for recognizing the orientation of a third monomer).

Let us now suppose that the monomers α and β have been correctly oriented and located, and that we want to orient the monomer γ. Then

$$F = F_\alpha + F_\beta + F_{c(\alpha,\beta)} = F_{\alpha,\beta} + F_{c(\alpha,\beta)},$$

where $F_{c(\alpha,\beta)}$ is the structure factor corresponding to the rest of the protein structure, and $\langle F\rangle = D_{\alpha,\beta}F_{\alpha,\beta}$, $\langle |F|^2\rangle = |F_{\alpha,\beta}|^2 + [\sum_N - \sum_{2N'/n}]$.

The structure factors R' and $R'_{\alpha,\beta}$ are now obtained by pseudo-normalizing F and $F_{\alpha,\beta}$ with respect to $[\sum_N - \sum_{2N'/n}]$. Under the above conditions it is easy to generalize the criterion (30) into

$$\textstyle\sum_i M_{ui}(\overline{E}_{\gamma i} - 1)(\langle R^2_{c(\alpha,\beta)i}\rangle - 1) = \max, \tag{31}$$

where

$$\langle R^2_{c(\alpha,\beta)i}\rangle = K[R^2_i + (\sigma^2_A)_{(\alpha,\beta)i} R^2_{(\alpha,\beta)i} - 2X_i m_1(2X_i)],$$

$$X_i = 2(\sigma_A)_{(\alpha,\beta)i} R_i R_{(\alpha,\beta)i}, \qquad K = \frac{\sum_N}{\sum_N - \sum_{2N'/n}}.$$

The generalization for finding the orientation of a fourth (or upper) monomer is trivial.

9. Translate a well oriented monomer when one or more monomers have been correctly oriented and located

Suppose that the monomer α has been correctly oriented and located, and that we want to locate the well oriented monomer β. Let us first suppose that the correct translation $\mathbf{N}_\beta$ has been found: then the model protein, constituted by the two monomers and their symmetry equivalents, will be an (imperfect) isomorph of the protein structure. In this case we define

$$\overline{F} = \overline{F}_\alpha + \sum_{s=1}^{m} a_{\beta,s,}\gamma_{\beta,s}, \qquad F = \sum_{s=1}^{m}\sum_{\mu=1}^{n} a_s g_{\mu,s},$$

$$\overline{E} = \overline{F} / \textstyle\sum^{1/2}_{2N'/n}, \qquad E = F / \sum^{1/2}_N,$$

where $\overline{F}_\alpha$ is fixed, and we can look for the joint probability distribution $P(E, \overline{E})$ under the following conditions:

(*a*) The coordinates of the vectors $\mathbf{r}'_j$, $j = 1, \ldots, p/n$, are fixed and *a priori* known; the coordinates of the vectors $\mathbf{r}'_j$, $j = p/n + 1, \ldots, 2p/n$, are primitive random variables, submitted to a geometric constraint (the geometry and the orientation of the model molecule is known).

(*b*) The variables $\mathbf{r}_j = \mathbf{r}'_j + \Delta\mathbf{r}_j$, $j = 1, \ldots, p/n$, are riding variables, correlated with the vectors $\mathbf{r}'_j$, $j = 1, \ldots, p/n$, through the local positional errors $\Delta\mathbf{r}_j$.

(*c*) The variables $\mathbf{r}_j = \mathbf{r}'_j + \Delta\mathbf{r}_j$, $j = p/n + 1, \ldots, 2p/n$, are riding variables. They are uncorrelated with the corresponding $\mathbf{r}_j$ owing to the fact that the position of the model molecule is unknown, but the interatomic distances $\mathbf{r}'_i - \mathbf{r}'_j$ are correlated with the vectors $\mathbf{r}_i - \mathbf{r}_j$ through the local positional errors $\Delta\mathbf{r}_j$.

(*d*) The variables $\mathbf{r}_j$, for $j = 2p/n + 1, \ldots, t$, are primitive random variables.

The probability distribution (23) may be used to derive

$$P(R, \overline{R}) \simeq 4R\overline{R}\frac{1}{(1-\sigma_A^2)}\exp\left[-\frac{1}{(1-\sigma_A^2)}(R^2+\overline{R}^2)\right]I_0(X), \tag{32}$$

where

$$\begin{aligned} X &= 2(\sigma_A)_{\alpha,\beta}R\overline{R}/\left[1-(\sigma_A^2)_{\alpha,\beta}\right], \\ (\sigma_A)_{\alpha,\beta} &= D_{\alpha,\beta}\left(\textstyle\sum_{2N'/n}/\sum_N\right)^{1/2} \qquad \text{if } 2N'/n \le N, \\ (\sigma_A)_{\alpha,\beta} &= D_{\alpha,\beta}\left(\textstyle\sum_N/\sum_{2N'/n}\right)^{1/2} \qquad \text{if } 2N'/n > N. \end{aligned}$$

$(\sigma_A)_{\alpha,\beta}$ and $D^2_{\alpha,\beta}$ are the values of σ_A and D^2 which may be obtained for the two-monomer case.

Let us now consider the case in which $\mathbf{N}_\beta$ does not coincide with the correct translation. Then R_β will be uncorrelated with R: accordingly $(\sigma_A)_{\alpha,\beta}$ will reflect the correlation between R and R_α only. The above considerations suggest that a useful criterion for finding the correct orientation is (see §7)

$$T = \textstyle\sum_i X_i m_1(X_i) = \max. \tag{33}$$

Let us examine the case in which the monomers α and β have been correctly oriented and located, and we look for the correct translation of the monomer γ. Let us first suppose that the correct translation N_γ has been found: then the model, constituted by the three monomers and their symmetry equivalents, is an (imperfect) isomorph of the protein structure, and

$$\overline{F} = \overline{F}_\alpha + \overline{F}_\beta + \sum_{s=1}^{m} a_{\gamma,s}, \gamma_{\gamma,s}, \qquad \overline{E} = \overline{F}/\textstyle\sum_{3N'/n}^{1/2}.$$

The same considerations made for the location of the monomer β suggest that (33) is still a useful criterion provided

$$\begin{aligned} X &= 2(\sigma_A)_{\alpha,\beta,\gamma}R\overline{R}/\left[1-(\sigma_A^2)_{\alpha,\beta,\gamma}\right], \\ (\sigma_A)_{\alpha,\beta,\gamma} &= D_{\alpha,\beta,\gamma}\left(\textstyle\sum_{3N'/n}/\sum_N\right)^{1/2} \qquad \text{if } 3N'/n < N, \\ (\sigma_A)_{\alpha,\beta,\gamma} &= D_{\alpha,\beta,\gamma}\left(\textstyle\sum_N/\sum_{3N'/n}\right)^{1/2} \qquad \text{if } 3N'/n > N. \end{aligned}$$

The procedure may be generalized as follows for any number of correctly oriented and translated monomers:

(*a*) The translation search for the second and upper monomers is made by fast Fourier transform as in *REMO*.

(*b*) The correct solution is found among the largest peaks *via* the criterion (33). The sum involves only the reflections for which $X > 1$. Of course, $\overline{F} = \overline{F}_\alpha + \overline{F}_\beta + \ldots + \sum_{s=1}^{m} a_{\varsigma,s}, \gamma_{\varsigma,s}$ and $\overline{R}$ is its normalized modulus. The monomer ς is the one we want to locate; the monomers $\alpha, \beta, \ldots$ are the monomers already located.

(*c*) For each reflection, $X = 2(\sigma_A)_{\alpha,\beta,\ldots,\varsigma}R\overline{R}/[1-(\sigma_A^2)_{\alpha,\beta,\ldots,\varsigma}]$.

As in *REMO*, the refinement of the position and of the orientation of each monomer is achieved by a subspace-searching *simplex* method (Rowan, 1990); it does not require the calculation of derivatives, but only the function evaluation.

10. Molecular replacement and pseudo-translational symmetry

Pseudo-translational symmetry is not rare in small-molecule or in protein crystallography: it may be ideal, or, more often, with strong deviations from ideality. Deviations may be (*a*) of replacive type, in this case the pseudo-translation vector $\mathbf{u}$ refers atoms of different species, or (*b*) of displacive type, when the corresponding atoms are slightly displaced from the ideal $\mathbf{u}$ vector. In real cases, replacive and displacive characters are simultaneously present. The presence and the nature of the pseudo-symmetry may be detected and characterized by a proper statistical analysis of diffraction data, just after the structure-factor normalization (Fan *et al.*, 1983; Böhme, 1982, 1983; Gramlich, 1984; Cascarano *et al.*, 1985, 1987, 1988). Information on $\mathbf{u}$, on the percentage of the electron density (percu) which satisfies the pseudo-translational symmetry, and on the nature of the pseudo-symmetry [if displacive, percu($|\boldsymbol{h}|$) is resolution dependent and diminishes at high $\sin\vartheta/\lambda$ values] can be actively used in MR to make the rotational and the translational search more fruitful.

In our treatment the pseudo-translational symmetry will be modelled according to Cascarano *et al.* (1988): we will consider only the case in which only one pseudo-translation vector $\mathbf{u}$ of order l is present. When $\langle\text{percu}\rangle$ [the average value of percu($|\boldsymbol{h}|$) for the resolution interval used in the MR calculations] is moderately large, one cannot assume that the related monomers have the same orientation: *e.g.* two monomers with centres of gravity related by $\mathbf{u}$ may have different orientations. The normalization section of *REMO09* automatically informs the user about the possible presence of pseudo-translational symmetry (*i.e.* when $\langle\text{percu}\rangle > 0.12$). The same feature is available on user demand in *MolRep* (Vagin & Teplyakov, 1997). If $\langle\text{percu}\rangle$ is sufficiently large it is reasonable to suppose that monomers related by $\mathbf{u}$ have also the same orientation.

11. Applications

In the theoretical part of this paper a number of mathematical criteria were obtained aiming at identifying the correct rotation and translation under different types of prior information: this section will check their efficiency. Instead of selecting a few specific cases, we preferred to apply *REMO09* to a large number of cases, listed in Table 1, to obtain statistically sound conclusions. The first 18 structures were used to test *REMO*: they represent either difficult cases for the authors who originally solved the crystal structures or test cases used by other authors to validate their MR programs. In the latter

Table 1
List of MR cases used to test the program *REMO09*.

Target is the PDB code of the protein structure; Model is the PDB code of the protein from which the model structure actually used has been extracted; NresT and NresM are the number of residues of the protein and model structures, respectively; RMS is the root-mean-square distance between the C_α atoms of model and target; Id is the sequence identity; *n* is the number of copies of the search model to locate by MR. Both RMS and Id are calculated for the aligned residues of the respective sequences. A reference is added when a test case was originally used to validate a program. The dashes for 1tgx in the column Model means: 'model by courtesy of Eleanor Dodson'.

Target / Model	NresT / NresM	*n*	RMS (Å)	Id (%)	Target / Model	NresT / NresM	*n*	RMS (Å)	Id (%)
1aki†‡/2ihl	129/129	1	0.39	97	1yxa/1qlp	740/372	2	1.68	46
1cgn†§/2ccy	125/127	1	1.73	31	2a03/1isc	394/384	1	0.89	53
1cgo†/2ccy	127/127	1	1.73	30	2a46/1g7k	217/217	1	0.95	40
1na7/1m2r	326/327	1	0.84	76	2a4k/1uls	439/245	2	1.08	64
1a6m/1mbc	151/153	1	0.22	100	2ah8/1ema	466/225	2	0.39	96
2iff¶/1hem	555/129	1	0.50	98	2ayv/1x23	148/153	1	0.79	80
6rhn†/4rhn	115/104	1	0.30	100	2b5o/1b2r	584/295	2	1.16	63
1kf3/7rsa	124/124	1	0.11	97	2f53/2bnr	811/820	1	1.03	98
1kqw/1opa	134/133	1	0.54	74	2f84/2aqw	323/321	1	0.99	67
1tp3/1be9	115/115	1	0.30	100	2f8m/1uj5	472/225	2	1.20	40
1bxo††/1er8	323/330	1	1.15	55	2fc3/1xbi	124/118	1	0.82	58
1zs0/1i76	165/163	1	0.36	100	2gq3/1n8i	1434/701	2	0.50	100
1e8a‡‡/1mho	175/88	2	1.52	36	2h8q/1g7k	868/218	4	0.33	97
2sar/1ulc	192/96	2	0.32	98	2hyu/1xjl	308/319	1	0.50	99
1lat§§/1glu	145/81	2	0.94	89	2hyw/1xjl	616/319	2	0.40	100
1lys‡/2ihl	258/129	2	0.60	96	2i3p/1g9y	304/304	1	0.33	99
6ebx/3ebx	124/62	2	0.82	100	2o3k/1ysb	307/317	1	0.35	99
1tgx/––––	180/50	3	0.65	100	2oka/20bk	336/335	1	0.45	87
1dy5/1lsq	248/124	2	0.27	100	2omt/1o6s	565/564	1	0.37	100
1s31/1c8z	273/265	1	1.26	96	2p0g/2oka	318/336	1	0.60	64
1xyg/1vkn	1380/1360	1	1.26	45	2qu5/2p2i	292/289	1	0.81	100
1ycn/1n00	619/318	2	1.16	72	2pby/1mki	1155/624	2	1.48	46

† *SOMoRe* (Jamrog *et al.*, 2003). ‡ *Queen of Spades* (Glykos & Kokkinidis, 2000). § *EPMR* (Kissinger *et al.*, 1999). ¶ *MolRep* (Vagin & Teplyakov, 1997). †† *Acon* (Yao, 2002). ‡‡ *MolRep* (Vagin & Teplyakov, 2000). §§ *Ultima* (Rabinovich *et al.*, 1998).

case, the programs involved are *SOMoRe* (Jamrog *et al.*, 2003), *Queen of Spades* (Glykos & Kokkinidis, 2000), *EPMR* (Kissinger *et al.*, 1999), *MolRep* (Vagin & Teplyakov, 1997), *Acorn* (Yao, 2002) and *Ultima* (Rabinovich *et al.*, 1998): the corresponding papers, which include the description of the original MR test, are referenced in Table 1. To enlarge the number of test cases, a further 26 structures, originally solved by MR, were randomly chosen among those deposited in the Protein Data Bank (PDB), for which diffraction data and information on the used models are available. The MR programs originally used to solve these structures are *EPMR*, *Amore* (Navaza, 1994) and *COMO* (Jogl *et al.*, 2001). In all the test cases the coordinate files of the model structures have been eventually modified according to the procedure originally used to solve the structure or to test MR programs.

In Table 1, the column Target/Model indicates the PDB code of the target and model structure, and the column NresT/NresM shows the number of residues of the protein and model structures. In particular, NresM refers to the part of the model actually used. A number of target residues very different from that of the model indicates that incomplete or overabundant models are used to fit the target. The column headed *n* shows the number of copies of the model structure to locate: $n > 1$ indicates that the target structure contains copies related by NCS.

To assess the difficulty of the various test cases, we used the *SSM* superposition algorithm (Krissinel & Henrick, 2004) included in the graphic tool *COOT* (Emsley & Cowtan, 2004): it provides two output parameters, both reported in Table 1; *i.e.* the root-mean-square distance between C_α atoms of search and target models (RMS) and the sequence identity (Id). They are both calculated for the aligned residues of the respective sequences, *i.e.* those residues which satisfy certain distance and orientation criteria at the best mutual superposition of the target and model structures. Therefore, a test case may be tagged as difficult if it has a high RMS value (then the backbone of the model significantly deviates from that of the target) and/or a low Id value (which indicates large differences between the target and model side chains).

The results of our tests may be summarized as follows.

(i) *Orient the first monomer*. In default conditions, for all the test structures the orientation with the highest score corresponds to the correct one, except for 1cgn and 1lat for which it is in position 4 and position 5, respectively. The first case is expected to be one of the most difficult: indeed it has the largest value of RMS (Å) and the smallest value of Id (%) among the full set of test structures; the second one has a high RMS value and is reported in the literature as solved only by including high-resolution reflections (up to 2 Å) in the calculations. Our results have been obtained without enhancing the contribution to R^2 arising from intramolecular vectors (avoiding procedures to separate them is advantageous when dealing with long and narrow models). In *REMO* and in other popular MR programs this is made by inverting the target Patterson map in a sphere of radius equal to the radius of gyration of the model structure.

If the correct orientation is searched utilizing all reflections, *i.e.* by including the reflections for which $R/2k_{200}$ and $\overline{E}_\beta$

values are close to unity, FOM_R slowly increases for the correct as well as for the incorrect solutions, leaving their contrast substantially unvaried. The order of the correct solution is maintained for all the structures except for 1cgo, for which the correct solution is lost.

(ii) *Orient a monomer when one or more other monomers have already been oriented*. The results are described in Table 2, where we give the score order for the correct solution (OO) and the number of combinations (Ncomb) among which the true orientation should be found. In nine of the 16 test cases the correct orientation is in position 1; in the other five cases the correct solutions are highly ranked. The worst cases are 1yxa and 2hyw: they are discussed below in this section. A third case deserves to be mentioned: for 1lat, in default conditions, OO/Ncomb is 13/136, obtained when the data-resolution limit is automatically fixed to 3.7 Å. On the contrary, OO is always 1 if the data-resolution limit is extended to better than 3 Å (1lat is reported in literature as solved only by including up to 2 Å resolution reflections).

(iii) *Translate a well oriented monomer*. The translation with the highest T score coincides with the correct one for all the test structures, except for 2hyw, 1dy5, 1yxa and 2pby, for which the correct translation is in position two: the first three of them are affected by pseudo-translational symmetry. The normalization routine of *REMO09* is able to verify the nature and the degree of pseudo-symmetry: this information may be actively employed by the user (see below).

(iv) *Orient a monomer when one or more other monomers have already been oriented and located*. The results are described in Table 2 (column OT/Ncomb), where we show the score order for the correct solution and the number of combinations among which the true orientation has been found. In 15 of the 17 test cases the correct orientation is in the first position; in the remaining cases the correct solution is in position 2 for 1e8a and position 8 for the first monomer of 2h8q. The largest effectiveness of the criteria (30) and (31) with respect to the criterion (20) suggested to us the following default procedure: the orientation of the second monomer is searched after having found the orientation and location of the first monomer, rather than by exploiting its orientation only.

(v) *Translate a well oriented monomer when one or more monomers have been correctly oriented and located*. The score order of the correct solution is always one.

(vi) *Rotate and translate when pseudo-translational symmetry is present*. Let us consider the following three cases, for which $n = 2$:

1dy5, with $\mathbf{u} = \mathbf{a}/2$, $\langle\mathrm{percu}\rangle = 0.51$,

2hyw, with $\mathbf{u} = (\mathbf{b} + \mathbf{c})/2$, $\langle\mathrm{percu}\rangle = 0.93$,

1lys, with $\mathbf{u} = (\mathbf{a} + \mathbf{c})/2$, $\langle\mathrm{percu}\rangle = 0.59$.

The substructure reflections are those for which $h = e$, $k + l = e$ and $h + l = e$, respectively.

Pseudo-translational symmetry can affect conventional estimates of figures of merit. Indeed, if we try to orient the second monomer of 1yxa and 2hyw given the orientation of the first (see §6) the correct orientations are both at position 16 (see Table 2, column OO/Ncomb). Similar effects are obtained for the translation search. Indeed, of the four structures (2hyw, 1dy5, 1yxa and 2pby) for which the correct translation vector is ranked at 2 by T, three (2hyw, 1dy5 and 1yxa) are affected by pseudo-translational symmetry.

Table 2
For each test structure with $n \geq 2$ (*a*) the heading OO shows the score order of the correct orientation for the second monomer when the orientation of the first is known; (*b*) OT gives the score order of the correct orientation of the second (and of the third, if the case) monomer when orientation and location of the first (and of the second) monomer are known; (*c*) Ncomb is the number of explored combinations in the two cases.

Target	n	OO/Ncomb	OT/Ncomb
1e8a	2	3/45	2/3
2sar	2	1/21	1/1
1lat	2	13/136	1/16
1lys	2	2/10	1/1
6ebx	2	2/15	1/1
1tgx	3	1/1	1/2, 1/1
1dy5	2	1/3	1/1
1ycn	2	1/1	1/1
1yxa	2	16/21	1/6
2a4k	2	1/21	1/1
2ah8	2	1/66	1/1
2b5o	2	1/15	1/1
2f8m	2	6/55	1/2
2gq3	2	1/1	1/1
2h8q	4	16/1365	8/14, 1/13, 1/12
2hyw	2	16/136	1/1
2pby	2	1/10	1/4

Pseudo-translational symmetry can also be considered a tool for a faster solution of the MR problem. Indeed, when it is strong, the second monomer can be directly oriented and located according to pseudo-translational information, so sparing computing time. Then a rigid-body refinement can improve orientation and translation parameters (see §12). For example, we applied the above technique to 1lys, 1dy5 and 2hyw, for which $\langle\mathrm{percu}\rangle > 0.50$, and we obtained final phase errors of 55°, 60° and 48°, respectively, comparable with 52°, 60° and 47° obtained by default.

An additional interesting case is 2p0g, for which $\mathbf{u} = \mathbf{c}/2$, $\langle\mathrm{percu}\rangle = 0.80$. The structure was originally solved by MR (Benach *et al.*, 2009), by using the four-monomer complex of the model structure 2oka. As stated in §2, we always reproduced, in our tests, the models originally used to solve the structures or to test MR programs: this is the reason why in Table 1 this case is characterized by the value $n = 1$. *REMO09* failed when it tried to solve the structure by using a model constituted by a single monomer: in fact, while the correct rotation was in position 1, the correct translation was discarded by the selection criterion. The translation step succeeded when we used the pseudo-translational information (*i.e.* by locating two equally oriented monomers separated by $\mathbf{u} = \mathbf{c}/2$): the correct translation was in position 1 and the other two monomers were easily located either by the procedure described in §7 or by exploiting the pseudo-translational vector as previously described.

A further example, 1yxa, characterized by a quite imperfect pseudo-translational symmetry, deserves to be discussed. The normalization section of *REMO09* did not find any appreciable pseudo-symmetry when the statistical analysis was performed at the same *RES* used for MR calculations (*i.e.* 4.25 Å). The same result is obtained utilizing *MolRep*, which, in this case, definitively loses the information (the correct pseudosymmetry is found by *MolRep* for 1dy5, 2hyw and 1lys). When we dropped *RES* from 4.25 to 6.81 Å, *REMO09* suggested $\mathbf{u} = \mathbf{a}/2$, $\langle\text{percu}\rangle = 0.23$, and a highly displacive deviation from ideality (*e.g.* $\langle\text{percu}\rangle = 0.80$ at very low resolution, about 0 for the resolution range close to 7 Å). To face problems like the above, we decided to include in the default procedure an additional statistical analysis for discovering hidden pseudo-translational symmetry: statistical calculations are first performed at the resolution chosen for solving the MR problem, then they are repeated by involving only lower resolution data (1.5 Å the previous resolution limit). The above considerations better explain the quite poor rank corresponding to the correct rotation of the second monomer given the rotation of the first (column OO/Ncomb in Table 2): this is due to the combination of pseudo-symmetry effects with lack of isomorphism between target and model molecule (RMS = 1.68 Å and Id = 46%).

12. Application of the DEDM-EDM procedure

The recovery of the target from the model structure usually requires two additional steps: (i) phase extension and refinement *via* EDM (electron density modification) techniques (Cowtan, 1999; Abrahams, 1997; Abrahams & Leslie, 1996; Zhang *et al.*, 2001; Refaat & Woolfson, 1993; Giacovazzo & Siliqi, 1997); (ii) electron density map interpretation in terms of the molecular model and its restrained refinement: manual inspection of the maps and/or automated model-building programs [*e.g. ARP/wARP* (Perrakis *et al.*, 1999), *PHENIX* (Terwilliger *et al.*, 2008), *MAID* (Levitt, 2001), *MAIN* (Turk, 2004), *Buccaneer* (Cowtan, 2006)] may be used.

Concerning point (i), we notice that EDM routines are usually external to MR programs. *REMO09* can automatically submit the MR solutions to the DEDM–EDM (difference electron density modification–electron density modification) approach, recently described by Caliandro *et al.* (2009*a*). The aim is to further reduce the phase error, so as to make the electron density map interpretation easier.

Concerning point (ii), *REMO09* electron density maps may be submitted to the DEA procedure (Caliandro *et al.*, 2009*b*) which uses an iterative combination of DEDM-EDM techniques with automated model-building programs (*i.e. PHENIX* and *ARP/wARP*). In this case *IL MILIONE* (the package in which *REMO09* is included) produces suitable scripts to manage the process.

13. Conclusions

A probabilistic model for molecular replacement has been described. The method uses the general approach of the joint probability distribution functions to establish criteria useful for identifying the orientation and the location of the monomers of the target structure by exploiting their structural similarity with the model monomers. The approach is also able to exploit various kinds of additional prior information which may be available at the different steps of the procedure. In particular, the previous knowledge of the orientation of one or more monomers is used to find better probabilistic criteria for orienting other monomers. Also, the orientation and the location of some monomers are used for the easier location of other well oriented monomers.

The probabilistic approach was implemented in the program *REMO09* and its efficiency was checked against a large set of test structures. The results are satisfactory: *REMO09* is able to provide, without any user intervention, useful electron density maps of the target structure, which may then be submitted to EDM-DEDM techniques for further phase extension and refinement. The program has been conceived to run with a high degree of automatism: diffraction data, expected cell content and model coordinates may be supplied in commonly used formats, through a user-friendly graphic interface. More different models may be put in, each one with a number of copies related by NCS. Modifications of the input models may be made through the graphic interface: new assignment of atomic thermal factors, cut of specific parts of the sequence and creation of polyalanine models. The program provides for an automatic selection of the reflections to be used during the MR search and of the feasible solutions. The user may modify the selections by putting in the sequence identity, or they may override them by inserting specific threshold values in the graphic window. A procedure to find pseudo-translational effects is performed in default: the user may decide whether to use it for locating copies of the given model.

At the end of the MR run, a packing calculation is performed on selected solutions. Trace atoms are identified in the model, by considering C_α atoms in the case of proteins: in the case of nucleic acids, phosphate and C atoms in the ribose-phosphate backbone and N atoms in the bases. The fraction of trace atoms that clash with their symmetry mates or with other trace atoms (symmetry mates included) is determined by considering 3.8 Å as the minimum allowed distance. Our tests suggest that *REMO09* is expected to be successful for MR cases which have Id > 30%, RMS < 1.75 Å and model completeness > 20%, the borderline test structures being 1cgno, 1cgn, 2iff, 1e8a, 1yxa, 2h8q and 2pby.

The program *REMO09* is included in release 2.2 of the package *IL MILIONE*, which is free for not-for-profit organizations and available for download, under License Agreement, from the site http://www.ic.cnr.it/.

APPENDIX *A*
Orienting the first monomer

Under the conditions specified in §4 for $\overline{F}$ and $\overline{E}$, and in §5 for E, we want to derive the characteristic function of the distribution $P(A, B, \overline{E})$. The lower-order cumulants k of such distribution are calculated below,

$$\langle \overline{F} \rangle = \sum_{s=1}^{m} \sum_{j=1}^{p/n} f_j'^2 = \textstyle\sum_{N'/n},$$

$$\langle \overline{E} \rangle = m_{001} = k_{001} = \langle \overline{F} \rangle / \textstyle\sum_{N'/n} = 1.$$

Since (see §4)

$$\langle \overline{E}^2 \rangle = m_{002} = (m+1)/m \neq 1,$$

$\overline{E}$ is a real non-negative pseudo-normalized function, while (see §5)

$$E = A + iB = F / \langle |F|^2 \rangle^{1/2} = F / \textstyle\sum_N^{1/2}$$

is a normalized structure factor. Let us now derive the value of $\langle |F|^2 \overline{F} \rangle$ when $\mathbf{M}_{\rm prot} = \mathbf{M}_{\rm mod}$,

$$\begin{aligned} \langle |F|^2 \overline{F} \rangle = \Big\langle & \sum_{s_1, s_2 = 1}^{m} \sum_{j_1, j_2 = 1}^{t} f_{j_1} f_{j_2} \exp\big[2\pi i \,\overline{\mathbf{h}}_{\rm prot}\big(\mathbf{R}_{s_1} \mathbf{r}_{j_1} - \mathbf{R}_{s_2} \mathbf{r}_{j_2} \\ & + \mathbf{T}_{s_1} - \mathbf{T}_{s_2}\big)\big] \\ & \times \sum_{s_3 = 1}^{m} \sum_{j_3, j_4 = 1}^{p/n} f_{j_3}' f_{j_4}' \exp\big[2\pi i \,\overline{\mathbf{h}}_{\rm prot} \mathbf{R}_{s_3} (\mathbf{r}_{j_3}' - \mathbf{r}_{j_4}')\big]\Big\rangle. \end{aligned} \qquad (34)$$

Non-vanishing contributions to the right-hand side of (34) arise as follows:

(*a*) For any value of s_3, when $s_1 = s_2$, $j_1 = j_2$, $j_3 = j_4$. The corresponding contribution is

$$\left(\sum_{s=1}^{m} \sum_{j=1}^{t} f_j^2 \right) \left(\sum_{s=1}^{m} \sum_{j_2=1}^{p/n} f_j'^2 \right) = \textstyle\sum_N \sum_{N'/n}, \qquad (35)$$

and corresponds to $\langle |F|^2 \rangle \langle \overline{F} \rangle$.

(*b*) For the atoms of the first monomer, when $s_1 = s_2 = s_3$, and, simultaneously, when $j_1 = j_4$, $j_2 = j_3$. The corresponding contribution is

$$\begin{aligned} & \Big\langle \sum_{s=1}^{m} \sum_{j_1, j_2 = 1}^{q} f_{j_1}^2 f_{j_2}'^2 \exp[2\pi i \,\overline{\mathbf{h}}_{\rm prot} \mathbf{R}_s (\Delta \mathbf{r}_{j_1} - \Delta \mathbf{r}_{j_2})] \Big\rangle \\ & = D^2 \sum_{s=1}^{m} \sum_{j_1, j_2 = 1}^{q} f_{j_1}^2 f_{j_2}'^2, \end{aligned} \qquad (36)$$

where $q = p/n$ if $p \le t$, $q = t/n$ if $p > t$. The above result has been obtained under the assumption (not entirely satisfied in real cases) that $\Delta \mathbf{r}_{j_1}$ and $\Delta \mathbf{r}_{j_2}$ are uncorrelated. Owing to the fact that most of the protein, as well as most of the model atoms, are light atoms (*i.e.* C, N, O) we can use the following approximation,

$$D^2 \sum_{s=1}^{m} \sum_{j_1, j_2 = 1}^{q} f_{j_1}^2 f_{j_2}'^2 \simeq D^2 m \textstyle\sum_q^2 .$$

Finally

$$\langle |F|^2 \overline{F} \rangle \simeq \textstyle\sum_N \sum_{N'/n} + D^2 m \sum_q^2 . \qquad (37)$$

In terms of normalized structure factors, (37) becomes

$$\langle |E|^2 \overline{E} \rangle \simeq \left(1 + \frac{D^2 m \sum_q^2}{\sum_N \sum_{N'/n}} \right).$$

To derive the value of m_{201} we observe that A and B play a symmetrical role in the product

$$\langle |E|^2 \overline{E} \rangle = \langle (A^2 + B^2) \overline{E} \rangle = 2 \langle A^2 \overline{E} \rangle = 2 m_{201},$$

and therefore

$$m_{201} = \frac{1}{2} \left(1 + \frac{D^2 m \sum_q^2}{\sum_N \sum_{N'/n}} \right).$$

The cumulants of the distribution are now easy found:

$$\begin{aligned} & k_{200} = m_{200} = k_{020} = m_{020} = 1/2, \\ & k_{002} = m_{002} - m_{001}^2 = 1/m, \\ & k_{201} = m_{201} - m_{200} m_{001} = m_{201} - 1/2 = \frac{1}{2} \frac{D^2 m \sum_q^2}{\sum_N \sum_{N'/n}}. \end{aligned}$$

It is easy to verify that if $p \le t$ then $m \sum_q^2 = m^{-1} \sum_{N'/n}^2$ and $k_{201} = (1/2m) D^2 (\sum_{N'/n} / \sum_N)$.

If $p > t$, then $m \sum_q^2 = m^{-1} \sum_{N/n}^2$ and $k_{201} = (1/2mn^2)$ $D^2 (\sum_N / \sum_{N'/n})$.

The characteristic function $C(u_1, v_1, u_{2,})$ of the distribution $P(A, B, \overline{E})$ is

$$\begin{aligned} C(u_1, v_1, u_{2,}) = {} & \exp\big\{-\tfrac{1}{4}\big[\big(u_1^2 + v_1^2\big) + iu_2 - (1/2m) u_2^2\big] \\ & + \tfrac{1}{2} k_{201} \big(u_1^2 u_2 + v_1^2 u_2\big)\big\}, \end{aligned} \qquad (38)$$

where u_1, v_1, u_2 are carrying variables associated with A, B and $\overline{E}$, respectively. The desired distribution is the Fourier transform of (38),

$$\begin{aligned} P(A, B, \overline{E}) = (2\pi)^{-3} & \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \exp\big\{-i[u_1 A + v_1 B + u_2 (\overline{E} - 1)]\big\} \\ & \times \exp\big[-\tfrac{1}{4}(u_1^2 + v_1^2) - (1/m) u_2^2\big] \\ & \times \big[1 + \tfrac{1}{2} k_{201} (u_1^2 u_2 + v_1^2 u_2)\big] \, {\rm d}u_1 \, {\rm d}v_1 \, {\rm d}u_2. \end{aligned}$$

We obtain

$$\begin{aligned} P(A, B, \overline{E}) = {} & 2^{-1/2} \pi^{-3/2} m^{1/2} \exp\big[-(A^2 + B^2 + \tfrac{1}{2} m (\overline{E} - 1)^2\big] \\ & \times \big\{1 + m k_{201} (\overline{E} - 1)\big[(2B^2 - 1) + (2A^2 - 1)\big]\big\}. \end{aligned} \qquad (39)$$

The change of variables $A = R \cos \phi$, $B = R \sin \phi$, leads to equation (12*a*).

APPENDIX *B*
Orienting a monomer when one or more others have already been oriented

In the following the lower moments of the distribution $P(A, B, \overline{E})$ under the assumptions defined in §6 are calculated. It may be useful to pseudo-normalize (instead of normalize) $\overline{F} = \overline{F}_\alpha + \overline{F}_\beta$ with respect to $\sum_{N'/n}$ and normalize F with respect to $\sum_N^{1/2}$. In this way the probabilistic formulas will be expressed in terms of the same $\overline{E}_\alpha$ and E factors employed for the search of the correct orientation of the α monomer. We obtain (see Appendix *A*):

$$\begin{aligned}
\overline{F}_{\alpha+\beta} &= \overline{F}_\alpha + \overline{F}_\beta,\\
\langle \overline{F}_{\alpha+\beta}\rangle &= \overline{F}_\alpha + \textstyle\sum_{N'/n},\\
\langle \overline{E}_{\alpha+\beta}\rangle &= 1 + \overline{F}_\alpha / \textstyle\sum_{N'/n} = \overline{E}_\alpha + 1,\\
\overline{F}^{\,2}_{\alpha+\beta} &= \overline{F}^{\,2}_\alpha + \overline{F}^{\,2}_\beta + 2\overline{F}_\alpha \overline{F}_\beta,\\
\langle \overline{F}^{\,2}_{\alpha+\beta}\rangle &= \overline{F}^{\,2}_\alpha + \frac{m+1}{m}\textstyle\sum^2_{N'/n} + 2\overline{F}_\alpha \textstyle\sum_{N'/n},\\
\langle \overline{E}^{\,2}_{\alpha+\beta}\rangle &= \overline{E}^{\,2}_\alpha + \frac{m+1}{m} + 2\overline{E}_\alpha = (\overline{E}_\alpha + 1)^2 + \frac{1}{m},
\end{aligned}$$

$$\langle \overline{E}^{\,2}_{\alpha+\beta}\rangle - \langle \overline{E}_{\alpha+\beta}\rangle^2 = m^{-1} = k_{002}.$$

Since $\langle \overline{E}^{\,2}\rangle \neq 1$, $\overline{E}$ is a pseudo-normalized variable.

For the variable F we have

$$\langle |F|^2\rangle = D^2_\alpha \overline{F}_\alpha + \textstyle\sum_{N(n-1)/n} = D^2_\alpha \overline{F}_\alpha + \dfrac{(n-1)}{n}\textstyle\sum_N, \qquad (40)$$

where D_α takes into account the misfit between the protein and the monomer α.

Let us define

$$E = A + iB = F/\textstyle\sum^{1/2}_N,$$

then

$$m_{100} = \langle A\rangle = m_{010} = \langle B\rangle = 0$$

$$\begin{aligned}
\langle A^2\rangle = m_{200} = k_{200} &= \frac{1}{2}\left[D^2_\alpha \overline{E}_\alpha + \frac{(n-1)}{n}\right]\\
&= \langle B^2\rangle = m_{020} = k_{020}.
\end{aligned}$$

Derive now the value of $\langle |F|^2 \overline{F}_{\alpha+\beta}\rangle$ when $\mathbf{M}_{\rm prot} = \mathbf{M}_{\beta\,{\rm mod}}$,

$$\langle |F|^2 \overline{F}_{\alpha+\beta}\rangle = \langle |F|^2 (\overline{F}_\alpha + \overline{F}_\beta)\rangle = \overline{F}_\alpha \langle |F|^2\rangle + \langle |F|^2 \overline{F}_\beta\rangle. \qquad (41)$$

In accordance with (40),

$$\overline{F}_\alpha \langle |F|^2\rangle = D^2_\alpha \overline{F}^{\,2}_\alpha + \frac{(n-1)}{n}\textstyle\sum_N \overline{F}_\alpha.$$

Derive now the value of $\langle |F|^2 \overline{F}_\beta\rangle$ when $\mathbf{M}_{\rm prot} = \mathbf{M}_{\beta\,{\rm mod}}$,

$$\begin{aligned}
\langle |F|^2 \overline{F}_\beta\rangle = \Bigg\langle \Bigg\{ D^2_\alpha \overline{F}_\alpha &+ \sum_{s_1,s_2=1}^{m}\ \sum_{j_1,j_2=t/n+1}^{t} f_{j_1} f_{j_2}\\
&\times \exp\big[2\pi i\, \overline{\mathbf{h}}_{\rm prot}(\mathbf{R}_{s_1}\mathbf{r}_{j_1} - \mathbf{R}_{s_2}\mathbf{r}_{j_2} + \mathbf{T}_{s_1} - \mathbf{T}_{s_2})\big]\Bigg\}\\
&\times \sum_{s_3=1}^{m}\ \sum_{j_3,j_4=1}^{p/n} f'_{j_3} f'_{j_4} \exp\big[2\pi i\, \overline{\mathbf{h}}_{\rm prot}\mathbf{R}_{s_3}(\mathbf{r}'_{j_3} - \mathbf{r}'_{j_4})\big]\Bigg\rangle \qquad (42)
\end{aligned}$$

Non-vanishing contributions to the right-hand side of (42) arise:

(*a*) for any value of s_3, when $s_1 = s_2$, $j_1 = j_2$, $j_3 = j_4$. The corresponding contribution is

$$\begin{aligned}
D^2_\alpha \overline{F}_\alpha \textstyle\sum_{N'/n} &+ \left(\displaystyle\sum_{s=1}^{m}\ \sum_{j=t/n+1}^{t} f^2_j\right)\left(\displaystyle\sum_{s=1}^{m}\ \sum_{j_2=1}^{p/n} f'^{\,2}_j\right)\\
&= D^2_\alpha \overline{F}_\alpha \textstyle\sum_{N'/n} + \dfrac{n-1}{n}\textstyle\sum_N \textstyle\sum_{N'/n}; \qquad (43)
\end{aligned}$$

(*b*) for the atoms of the first monomer, when $s_1 = s_2 = s_3$, and, simultaneously, when $j_1 = j_4$, $j_2 = j_3$. The corresponding contribution is

$$\begin{aligned}
&\left\langle \sum_{s=1}^{m}\ \sum_{j_1,j_2=1}^{q} f^2_{j_1} f'^{\,2}_{j_2} \exp[2\pi i\, \overline{\mathbf{h}}_{\rm prot}\mathbf{R}_s(\Delta\mathbf{r}_{j_1} - \Delta\mathbf{r}_{j_2})]\right\rangle\\
&\quad = D^2_\beta \sum_{s=1}^{m}\ \sum_{j_1,j_2=1}^{q} f^2_{j_1} f'^{\,2}_{j_2} \simeq D^2_\beta\, m \textstyle\sum^2_q \qquad (44)
\end{aligned}$$

where $q = p/n$ if $p \le t$, $q = t/n$ if $p > t$, and D_β is the value of D calculated for the monomer β. Therefore

$$\langle |F|^2 \overline{F}_\beta\rangle = D^2_\alpha \overline{F}_\alpha + \frac{n-1}{n}\textstyle\sum_N \textstyle\sum_{N'/n} + mD^2_\beta \textstyle\sum^2_q .$$

Then

$$\begin{aligned}
&\langle |F|^2 \overline{F}_{\alpha+\beta}\rangle - \langle |F|^2\rangle\langle \overline{F}_{\alpha+\beta}\rangle\\
&\quad = D^2_\alpha \overline{F}^{\,2}_\alpha + \frac{(n-1)}{n}\textstyle\sum_N \overline{F}_\alpha + D^2_\alpha \overline{F}_\alpha \textstyle\sum_{N'/n}\\
&\quad\quad + \frac{n-1}{n}\textstyle\sum_N \textstyle\sum_{N'/n} + mD^2_\beta \textstyle\sum^2_q\\
&\quad\quad - \left[D^2_\alpha \overline{F}_\alpha + \frac{(n-1)}{n}\textstyle\sum_N\right]\left(\overline{F}_\alpha + \textstyle\sum_{N'/n}\right)\\
&\quad\quad = mD^2_\beta \textstyle\sum^2_q .
\end{aligned}$$

Owing to the symmetrical role played by A and B in the average $\langle |F|^2 \overline{F}\rangle$ we have

$$m_{201} = m_{021} = 0.5\langle |E|^2 \overline{E}\rangle;$$

accordingly

$$\begin{aligned}
k_{201} &= \frac{1}{2m} D^2_\beta \frac{\sum_{N'/n}}{\sum_N} \qquad \text{if } p \le t \quad \text{and}\\
k_{201} &= \frac{1}{2mn^2} D^2_\beta \frac{\sum_N}{\sum_{N'/n}} \qquad \text{if } p > t.
\end{aligned}$$

The characteristic function of $P(A, B, \overline{E})$ is

$$\begin{aligned}
C(u_1, v_1, u_2,) = {} & \exp\big\{-\tfrac{1}{2}\big[k_{200}u^2_1 + k_{020}v^2_1\big]\\
& + i(\overline{E}_\alpha + 1)u_2 - (1/2m)u^2_2\\
& + \tfrac{1}{2}k_{201}(u^2_1 u_2 + v^2_1 u_2)\big\}, \qquad (45)
\end{aligned}$$

where u_1, v_1, u_2 are carrying variables associated with $A, B, \overline{E}$, respectively.

Equation (45) is formally equivalent to (38) if the following variable changes are introduced,

$$\begin{aligned}
u_1 = u'_1(2k_{200})^{-1/2}, \quad v_1 = v'_1(2k_{020})^{-1/2}, \quad u_2 = u'_2(\overline{E}_\alpha + 1)^{-1},\\
m' = m(\overline{E}_\alpha + 1)^2, \quad k'_{201} = k_{201}/[2k_{200}(\overline{E}_\alpha + 1)].
\end{aligned}$$

Then its Fourier transform may be calculated according to (39). We first obtain $P(A', B', \overline{E}')$, where $A' = A(2k_{200})^{-1/2}$, $B' = B(2k_{020})^{-1/2}$, $\overline{E}' = \overline{E}(\overline{E}_\alpha + 1)^{-1}$, and then $P(A, B, \overline{E})$,

$$\begin{aligned}
P(A, B, \overline{E}) \simeq S\exp&\left[-\frac{(A^2 + B^2)}{2k_{200}} - \frac{1}{2}m(\overline{E} - \overline{E}_\alpha - 1)^2\right]\\
&\times \left\{1 + m\frac{k_{201}}{2k_{200}}(\overline{E} - \overline{E}_\alpha - 1)\left[\frac{(A^2 + B^2)}{(2k_{200})} - 1\right]\right\},
\end{aligned} \qquad (46)$$

where S is a suitable normalization factor. The change of variables $A = R\cos\phi$, $B = R\sin\phi$ leads to equation (15).

APPENDIX *C*
Orient a monomer when one or more other monomers have already been oriented and located

Here the normalization process and the subsequent probabilistic calculations will be performed in accordance with the conditions stated in §8. In particular,

(*a*) Since $\langle\overline{F}\rangle = \sum_{s=1}^{m}\sum_{j=1}^{p/n} f_j'^2 = \sum_{N'/n}$, then (see §4)

$$\langle\overline{E}\rangle = \langle\overline{F}\rangle/\textstyle\sum_{N'/n} = 1, \qquad \langle\overline{F}^2\rangle = \dfrac{m+1}{m}\textstyle\sum_{N'/n}^2,$$
$$\langle\overline{E}^2\rangle = (m+1)/m \neq 1.$$

$\overline{E}$ is a real non-negative pseudo-normalized quantity.

(*b*) Since

$$\langle F\rangle = D_\alpha F_\alpha, \quad \langle|F|^2\rangle = |F_\alpha|^2 + \left(\textstyle\sum_N - \sum_{N'/n}\right),$$

we pseudo-normalize F and F_α with respect to the unlocated electron density,

$$E' = (A' + iB') = F/\left(\textstyle\sum_N - \sum_{N'/n}\right)^{1/2},$$
$$E'_\alpha = (A'_\alpha + iB'_\alpha) = F_\alpha/\left(\textstyle\sum_N - \sum_{N'/n}\right)^{1/2}.$$

The first moments of the distribution $P(A', B', \overline{E})$ are

$$m_{100} = k_{100} = D_\alpha A'_\alpha, \quad m_{010} = k_{010} = D_\alpha B'_\alpha,$$
$$m_{001} = k_{001} = 1, \qquad m_{200} = A'^2_\alpha + 0.5, \quad m_{020} = B'^2_\alpha + 0.5,$$
$$k_{200} = k_{020} = 0.5, \quad m_{002} = (m+1)/m, \quad k_{002} = m^{-1}.$$

In accordance with Appendices *A* and *B*,

$$\langle|F|^2\overline{F}\rangle = \langle|F_\alpha + F_{c\alpha}|^2\overline{F}\rangle = |F_\alpha|^2\langle\overline{F}\rangle + \langle|F_{c\alpha}|^2\overline{F}\rangle$$

and

$$\langle|F|^2\overline{F}\rangle - \langle|F|^2\rangle\langle\overline{F}\rangle = D_\beta^2 m \textstyle\sum_q^2,$$

where $q = p/n$ if $p > t$, $q = t/n$ if $p > t$, and D_β is the value of D calculated for the monomer β. Owing to the symmetrical role played by A and B in the average $\langle|F|^2\overline{F}\rangle$ we have

$$m_{201} = m_{021} = 0.5\langle|E|^2\overline{E}\rangle,$$
$$k_{201} = \frac{1}{2m}D_\beta^2\frac{\sum_{N'/n}}{\sum_N} \qquad \text{if } p \leq t,$$
$$k_{201} = \frac{1}{2mn^2}D_\beta^2\frac{\sum_N}{\sum_{N'/n}} \qquad \text{if } p < t.$$

The characteristic function $C(u_1, v_1, u_{2,})$ of the distribution $P(A', B', \overline{E})$ is

$$C(u_1, v_1, u_{2,}) = \exp\Big[i(D_\alpha A'_\alpha u_1 + D_\alpha B'_\alpha v_1 + u_2) - \tfrac{1}{4}(u_1^2 + v_1^2) - (1/2m)u_2^2 + \tfrac{1}{2}k_{201}(u_1^2 u_2 + v_1^2 u_2)\Big],$$

from which

$$P(A', B', \overline{E}) \simeq (2\pi)^{-3/2}2m^{1/2}\exp\Big\{-\Big[(A' - D_\alpha A'_\alpha)^2 + (B' - D_\alpha B'_\alpha)^2\Big] - m(\overline{E} - 1)\Big\} \times \Big\{1 + 2k_{201}m(\overline{E} - 1)\Big[(A' - D_\alpha A'_\alpha)^2 - 0.5 + (B' - D_\alpha B'_\alpha)^2 - 0.5)\Big]\Big\}.$$

The change of variables $A' = R'\cos\phi$, $B' = R'\sin\phi$ leads to

$$P(R', \phi, \overline{E}) = (2\pi)^{-3/2}2m^{1/2}\exp\Big\{-\Big[R'^2 + D_\alpha^2 R'^2_\alpha - 2D_\alpha R' R'_\alpha\cos(\phi - \phi_\alpha)\Big] - m(\overline{E} - 1)\Big\} \times \Big\{1 + 2k_{201}m(\overline{E} - 1)\Big[R'^2 + D_\alpha^2 R'^2_\alpha - 2D_\alpha R' R'_\alpha\cos(\phi - \phi_\alpha) - 1\Big]\Big\}.$$

The integration over ϕ leads to equation (28).

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Benach, J., Neely, H., Seetharaman, J., Ho, C. K., Janjua, H., Cunningham, K., Ma, L., Xiao, R., Liu, J., Baran, M. C., Acton, T. B., Rost, B., Montelione, G. T., Hunt, J. F. & Tong, L. (2009). In preparation.

Böhme, R. (1982). *Acta Cryst.* A**38**, 318–326.

Böhme, R. (1983). *Z. Naturforsch.* **38**, 304–307.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G., Siliqi, D. & Spagna, R. (2007). *J. Appl. Cryst.* **40**, 609–613.

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., Giacovazzo, C., Mazzone, A. M. & Siliqi, D. (2009*a*). *Acta Cryst.* D**65**, 249–256.

Caliandro, R., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mazzone, A. M. & Siliqi, D. (2009*b*). *Acta Cryst.* D**65**, 477–484.

Carrozzini, B., Cascarano, G. L. & Giacovazzo, C. (2009). *J. Appl. Cryst.* Submitted.

Cascarano, G., Giacovazzo, C. & Luić, M. (1985). *Acta Cryst.* A**41**, 544–551.

Cascarano, G., Giacovazzo, C. & Luić, M. (1987). *Acta Cryst.* A**43**, 14–22.

Cascarano, G., Giacovazzo, C. & Luić, M. (1988). *Acta Cryst.* A**44**, 183–189.

Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.

Cowtan, K. (1999). *Acta Cryst.* D**55**, 1555–1567.

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

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Fan, Hai-Fu, Yao, Jia-Xing, Main, P. & Woolfson, M. M. (1983). *Acta Cryst.* A**39**, 566–569.

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

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

Gramlich, V. (1984). *Acta Cryst.* A**40**, 610–616.

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

Jogl, G., Tao, X., Xu, Y. & Tong, L. (2001). *Acta Cryst.* D**57**, 1127–1134.

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

Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.

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

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

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.

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

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.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.

Read, R. J. (1999). *Acta Cryst.* D**55**, 1759–1764.

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

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

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* A**46**, 783–792.

Rowan, T. (1990). PhD thesis, University of Texas at Austin, USA.

Sheriff, S., Klei, H. E. & Davis, M. E. (1999). *J. Appl. Cryst.* **32**, 98–101.

Shmueli, U. & Wilson, A. J. C. (1993). *International Tables for Crystallography*, Vol. B., pp. 184–200. Dordrecht: Kluwer Academic Publishers.

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

Srinivasan, R. & Subramanian, E. (1964). *Acta Cryst.* **17**, 67–68.

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.

Turk, D. (2004). *Acta Cryst.* A**60**, s16.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.

Vagin, A. & Teplyakov, A. (2000). *Acta Cryst.* D**56**, 1622–1624.

Yao, J.-X. (2002). *Acta Cryst.* D**58**, 1941–1947.

Zhang, K. Y. J., Cowtan, K. D. & Main, P. (2001). *International Tables for Crystallography*, Vol. F, pp. 311–331. Dordrecht: Kluwer Academic Publishers.