Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Refinement of severely incomplete structures with maximum likelihood in *BUSTER–TNT*

E. Blanc,[a,b]‡ P. Roversi,[a,c]*‡ C. Vonrhein,[a] C. Flensburg,[a] S. M. Lea[c] and G. Bricogne[a]

[a]Global Phasing Ltd, Sheraton House, Castle Park, Cambridge CB3 0AX, England, [b]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England, and [c]Laboratory of Molecular Biophysics, Department of Biochemistry, Oxford University, South Parks Road, Oxford OX1 3QU, England

‡ These authors contributed equally to this work.

Correspondence e-mail: pietro@biop.ox.ac.uk

Received 3 February 2004
Accepted 6 July 2004

BUSTER–TNT is a maximum-likelihood macromolecular refinement package. *BUSTER* assembles the structural model, scales observed and calculated structure-factor amplitudes and computes the model likelihood, whilst *TNT* handles the stereochemistry and NCS restraints/constraints and shifts the atomic coordinates, *B* factors and occupancies. In real space, in addition to the traditional atomic and bulk-solvent models, *BUSTER* models the parts of the structure for which an atomic model is not yet available ('missing structure') as low-resolution probability distributions for the random positions of the missing atoms. In reciprocal space, the *BUSTER* structure-factor distribution in the complex plane is a two-dimensional Gaussian centred around the structure factor calculated from the atomic, bulk-solvent and missing-structure models. The errors associated with these three structural components are added to compute the overall spread of the Gaussian. When the atomic model is very incomplete, modelling of the missing structure and the consistency of the *BUSTER* statistical model help structure building and completion because (i) the accuracy of the overall scale factors is increased, (ii) the bias affecting atomic model refinement is reduced by accounting for some of the scattering from the missing structure, (iii) the addition of a spatial definition to the source of incompleteness improves on traditional Luzzati and σ_A-based error models and (iv) the program can perform selective density modification in the regions of unbuilt structure alone.

1. Introduction

BUSTER–TNT (Bricogne & Irwin, 1996) is a maximum-likelihood macromolecular refinement package. *BUSTER* (Bricogne, 1993*a*) assembles the structural model, scales observed and calculated structure-factor amplitudes and computes the model likelihood. The structural model in *BUSTER* can include a description of the parts of the structure for which an atomic model is not yet available ('missing structure'). *TNT* (Tronrud *et al.*, 1987; Tronrud, 1992, 1996, 1997, 1999) receives the likelihood derivatives from *BUSTER*, evaluates the stereochemistry and NCS restraints residuals and their derivatives and shifts the coordinates, *B* factors and occupancies of the atomic model to maximize their likelihood while satisfying the restraints.

Both the use of maximum likelihood (ML) and the modelling of the missing structure help in overcoming the major drawbacks encountered by classical methods [least squares (LS) + difference maps] when dealing with the refinement and completion of incomplete structures.

(i) Recourse to ML instead of LS helps reduce overfitting of the observed amplitudes at phases too close to those of the

initial partial structure, by keeping an appropriate distance from the data. In the presence of non-negligible errors in the model, LS refinement is known to produce biased results in which the corrections to the initial partial structure are smaller than they ought to be (Bricogne & Irwin, 1996; Pannu & Read, 1996; Murshudov *et al.*, 1997, and references therein).

(ii) The atoms of the missing structure that contribute to the observed scattering are modelled using a real-space low-resolution probability distribution for their random positions (Bricogne, 1984, 1997; Roversi *et al.*, 2000), computed from any prior low-resolution information already available as to their placement. This process introduces a spatial non-uniformity to the source of incompleteness, which departs from the traditional uniform Wilson distribution of missing atoms encoded in σ_A-based error models. This non-uniformity in turn helps both the scaling process and the partial structure refinement.

The ML method and the missing-structure parameterization are based on a statistical treatment of model structure factors by techniques that constitute the core of *BUSTER* (Bricogne, 1988, 1993*a*). Their purpose is to generate and exploit quantitative descriptions of the statistical behaviour of structure factors resulting from the two main sources of randomness present in the situation described above:

(i) errors in the current atomic model, *i.e.* the imperfection of the atomic model;

(ii) uncertainty arising from the fact that the atoms that are missing from the atomic model cannot be represented by definite atomic parameters and must be treated as statistically distributed, *i.e.* the incompleteness of the atomic model.

At any given stage of the refinement or completion process, model structure factors do not have a 'calculated value' as implied by the usual notation $\mathbf{F}_{\rm calc}$; instead, they have a probability distribution. In practice, these distributions are often approximated by Gaussians and are hence described in terms of the expectation of any collection of random structure factors and by the covariance matrix of fluctuations around these expectations (Bricogne, 1988).

This statistical picture takes into account the phase uncertainty present in these model structure factors to drive the refinement of the partial structure. Instead of treating their phases as constants when trying to improve the fit between the model amplitudes and the observed ones, *BUSTER* calculates the marginal probability distribution of model amplitudes and seeks to maximize the value taken by this marginal probability over the observed amplitudes. This value is called the likelihood of the current model, Λ, and its maximization with respect to all or any of the parameters describing the current model is called the ML refinement of those parameters.

Unlike the LS method, the initial probability distribution for the model structure factors may contain an explicit dependency on parameters that influence the variance of the distribution and such parameters may be refined along with others. These parameters are referred to as imperfection parameters. It is through such refineable variance-modulating parameters that the ML method is able to keep a safe distance between observed amplitudes and the amplitudes of the traditional $\mathbf{F}_{\rm calc}$s and thus avoid overfitting. Experimental information on the phases attached to the observed amplitudes can further assist in this bias removal.

The ML refinement of the atomic model (in conjunction with *TNT*) and the parameterization of the missing structure by means of a low-resolution real-space distribution are naturally associated in this formalism in the sense that the probability distribution of the model structure factors and hence the likelihood Λ of the current model depends symmetrically on the atomic parameters (x, y, z, B, occupancies) describing the current partial structure and on other parameters, the Lagrange multipliers (λs), describing the extra detail currently conveyed by the positional distribution of the atoms in the missing structure. Since the model structure factors are sums of contributions from the partial structure, the missing structure and the bulk solvent, we see that the gradient of the log-likelihood (LL), $\mathcal{L} = \log(\Lambda)$, with respect to the expectations of model structure factors can be redirected (by the chain rule) either towards the atomic parameters on which the atomic model contribution depends, or towards the Lagrange multipliers on which the missing structure contribution depends, or towards both.

The present paper focuses on the partial structure refinement in *BUSTER–TNT*, while a manuscript in preparation will describe the phase refinement by ML variation of the missing atoms' λs in *BUSTER*.

2. Symbols used in this paper

Four types of real-space distributions are dealt with, all of which are handled in *BUSTER* as *CCP*4-format maps sampled on a crystallographic grid with N_x, N_y and N_z points along the crystallographic axes. We list here the symbols for these distributions (but omitting any subscripts) as an aid to the reader.

$f(\mathbf{x})$: a generic distribution in the crystallographic unit cell.

$q(\mathbf{x})$, $m(\mathbf{x})$: everywhere non-negative and continuous functions, normalized so that their average in the unit cell is unity,

$$(1/V)\int_V q(\mathbf{x})\,{\rm d}\mathbf{x} = 1, \tag{1}$$

V being the volume of the unit cell; when sampling $q(\mathbf{x})$ on a grid,

$$\frac{1}{N_x N_y N_z}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\sum_{k=1}^{N_z} q(i,j,k) = 1. \tag{2}$$

$\rho(\mathbf{x})$: an electron density, in e Å^{-3} units.

Other symbols are as follows.

Angle brackets: expectation value under a probability density, $\langle f\rangle = \int P(x) f(x)\,{\rm d}x$.

Angle brackets with a resolution suffix: average within a resolution bin,

$$\langle f\rangle_{d^*} = \frac{1}{N}\sum_{\mathbf{h}\in D^*}^{N} f(d^*_{\mathbf{h}}), \tag{3}$$

with $D^* = \{\mathbf{h} \mid d^* - \Delta^* \leq d^*_{\mathbf{h}} < d^* + \Delta^*\}$ and Δ^* the half-width of that resolution bin.

For the Fourier operators we follow the notation introduced in §1.3.3.1 of Bricogne (1993*b*).

3. The *BUSTER–TNT* structural model

The structural model used for the distribution of the atoms in the crystal in *BUSTER* has three components (or channels; Bricogne, 1988).[1]

(i) The partial structure (or fragment), the set of atoms for which positional coordinates, *B* factors and occupancies are available and can be refined.

(ii) The missing structure, defined as the part of the structure that is scattering in an orderly manner but has not yet been modelled with atomic positions and *B* factors.

(iii) The bulk solvent, defined as the disordered solvent atoms occupying the regions left void by the macromolecule in the crystal.

The distribution of the atoms in the crystal is written as

$$\rho^{\rm calc}(\mathbf{x}) = \rho^{\rm frag}(\mathbf{x}) + \rho^{\rm miss}(\mathbf{x}) + \rho^{\rm solv}(\mathbf{x}). \quad (4)$$

The linearity of the Fourier operator gives a sum of three terms for the total structure factor as well,

$$\mathbf{F}_{\mathbf{h}} = \mathbf{F}^{\rm frag}_{\mathbf{h}} + \mathbf{F}^{\rm miss}_{\mathbf{h}} + \mathbf{F}^{\rm solv}_{\mathbf{h}}. \quad (5)$$

The three individual components of the structure factor are all treated as random vectors, each of which is distributed as a two-dimensional Gaussian in the complex plane, carrying its own uncertainty model and variance (see §§3.1.1, 3.2.4 and 3.3.1).

Under the hypothesis that the three structural components are independent, their sum, the structure factor for the whole structure, is also treated as a random vector and is distributed in the complex plane according to a Gaussian, which is the product of the Gaussians for the three individual components.

The assumption that the errors of the partial structure and the bulk solvent are independent breaks down at low resolution, given that the Babinet opposite of the bulk-solvent envelope is most often computed by masking around the partial structure. Therefore, the error model at low resolution can be overly pessimistic, because the sum of variances would need a negative covariance to diminish the total variance.

3.1. The partial structure

The partial structure or fragment is the set of atoms for which positional coordinates, *B* factors and occupancies are available and can be refined. The electron density computed from this atomic model is denoted by $\rho^{\rm frag}(\mathbf{x})$.

3.1.1. The Luzzati distribution of $\mathbf{F}^{\rm frag}_{\mathbf{h}}$. We make the assumption that the distribution of each and every partial structure atom is a Gaussian centred around the mean atomic position and that the partial structure atoms are all distributed independently of one another. It is also assumed that *B* factors and occupancy values have no errors associated with them, in the sense that they follow a degenerate probability distribution with zero variance (Luzzati, 1952). The same Luzzati model can be used to model errors in the placement of a rigid-body 'fragment'.

Under these hypotheses for the distributions of positions, *B* factors and occupancies, the structure factor for the partial structure, $\mathbf{F}^{\rm frag}_{\mathbf{h}}$, is distributed around the offset $\mathbf{F}^{\rm frag,calc}_{\mathbf{h}}$ with a variance $\mathbf{V}^{\rm frag}_{\mathbf{h}}$ following a two-dimensional Gaussian whose first and second moments (or offset and variance) are computed as follows (Bricogne & Irwin, 1996).

(i) The offset is obtained from the partial structure model,

$$\mathbf{F}^{\rm frag,calc}_{\mathbf{h}} = \bar{\mathcal{F}}[\,\rho^{\rm frag}(\mathbf{x})](\mathbf{h})\, D^{\rm frag}_{\mathbf{h}}, \quad (6)$$

with $D^{\rm frag}_{\mathbf{h}} = \exp[-(1/4)B^{\rm frag}_{\rm impf}\, d^{*2}_{\mathbf{h}}]$. $B^{\rm frag}_{\rm impf}$ is the parameter modelling the imperfection of the partial structure. $\mathbf{F}^{\rm frag,calc}_{\mathbf{h}}$ is also called the 'attenuated fragment structure factor'.

(ii) The variance tensor, $\mathbf{V}^{\rm frag}_{\mathbf{h}}$, depends on the imperfection parameter $B^{\rm frag}_{\rm impf}$ and is diagonal; the first component, $\mathbf{V}^{11}_{\rm frag}$, refers to the real part and the second, $\mathbf{V}^{22}_{\rm frag}$, to the imaginary part of the structure factor. For each reflection $\mathbf{h}$, both diagonal elements, $\mathbf{V}^{ii}_{\rm frag}$, are set so that they are equal to half of the average partial structure intensity within the resolution bin for that reflection,

$${}^{\mathbf{h}}\mathbf{V}^{ii}_{\rm frag} = (1/2)\,\varepsilon_{\mathbf{h}}\langle\bar{\mathcal{F}}[\rho^{\rm frag}(\mathbf{x})]^2(\mathbf{h})\rangle_{d^*}\,[1 - (D^{\rm frag}_{\mathbf{h}})^2]. \quad (7)$$

When the fragment coordinate error is very small, the imperfection factor tends towards zero, the partial structure offset tends to the unattenuated fragment structure factor and the associated variance shrinks towards zero. In this limiting case, and provided that the model contains no other source of error, the present formalism tends towards a standard LS problem.

At the opposite end of the imperfection regime, with large coordinate errors, $B^{\rm frag}_{\rm impf}$ tends to infinity, the offset tends towards zero and the variance is the full fragment intensity in the resolution bin. The imperfection factor erases all previous knowledge of atom localization and the only remaining information comes from the number, type and temperature factor of the pool of missing atoms. This is the Wilson regime.

3.2. The missing structure

Within *BUSTER*, the atoms in the missing structure are described by adopting the random scatterer model, introduced to crystallography in the context of direct methods (Bricogne, 1984). According to this model, the missing atoms are all equally and independently distributed at random, following the real-space distribution $q^{\rm miss}(\mathbf{x})$, which can be modulated by maximum-likelihood refinement with entropy restraints (Bricogne, 1997; details of the modulation of $q^{\rm miss}$ will be given by a paper in preparation).

During the *BUSTER–TNT* refinement of the partial structure atomic model no modulation of $q^{\rm miss}$ is carried out and for all practical purposes this distribution can be thought of as a

[1] The first and third contribution to the crystal electron density are the familiar contributions present in all macromolecular refinement programs, while the second is to date present in *BUSTER* only.

prior probability distribution for the random positions of the atoms of the missing structure. In this case, $q^{\rm miss}$ is indicated as $m^{\rm miss}(\mathbf{x})$ or simply $m(\mathbf{x})$. With this probability model one can compute not only a value for the expectation of the low-resolution electron density for the missing structure but also a statistical variance around that expectation (the latter variance captured in reciprocal space, see §3.2.4).

The calculation of $m^{\rm miss}(\mathbf{x})$ is described in the next sections. Similar techniques can be used to compute the envelopes for the whole macromolecule or for the bulk solvent. A more detailed description of the algorithms described here is given by Roversi *et al.* (2000).

3.2.1. Uniform prior $m^{\rm miss}(\mathbf{x})$. The simplest choice for the prior probability distribution of the atoms in the missing structure is to exclude them from the regions that already contain a reliable atomic model; this approach brings into the statistical model the notion that a number of atoms are missing and that they are equally likely to be anywhere except where other atoms have been placed already.

The uniform prior distribution is defined in four steps.

(i) A binary mask is drawn around the atomic model.

(ii) This mask is symmetry-expanded to cover the whole cell.

(iii) The mask is negated to obtain a mask over the missing-structure region.

(iv) The mask is blurred by means of a convolution with an isotropic Gaussian and normalized. The convolution is carried out in reciprocal space, using a set of periodized ('aliased') structure factors (Roversi *et al.*, 1998) for $m^{\rm miss}(\mathbf{x})$, to ensure that the final distribution is everywhere non-negative and free from Fourier-truncation artefacts.

We stress that this distribution is uniform outside the regions occupied by the model, hence the name 'uniform prior', but its shape is not uniform; only in the absence of any atomic model would this be a truly uniform distribution throughout the unit cell.

We also notice that if the bulk-solvent envelope is chosen to fill up all the space left empty by the macromolecular model, the missing-structure envelope and the bulk-solvent envelope overlap. Although they can still differ for the parameter B used in the blurring step, this overlap introduces very large correlations between the scaling parameters for these two components.

3.2.2. Model-based non-uniform prior $m(\mathbf{x})$. Sometimes a rough guess is available as to the placement of a subset of atoms, such as a protein loop or domain or a bound ligand, but the model tentatively built for the same atoms is questionable.

An envelope $m^{\rm miss}(\mathbf{x})$ can then be built around these ill-defined atoms and the same atoms can be omitted from the partial structure. The real-space picture of the crystal in this case then comprises the bulk-solvent envelope, the atomic model for the trusted traced atoms and the missing-structure envelope. The latter is localized around the tentatively placed atoms; it represents our prior expectation about their position but does not retain any of the high-resolution details that are being assessed.

The prior distribution is computed in the same way as in the uniform prior case, except for the definition of the initial binary mask; a mask is built around the total structure (fragment and missing structure), from which a mask around the fragment alone is then subtracted. By suitably assigning the masking radii, this protocol allows for the generation of an envelope around a missing loop or domain or a layer of partially ordered solvent around the fragment.

Again, depending on the PDB models, masking radii and blurring factors used to compute the missing structure and bulk-solvent distributions, their boundaries can sometimes overlap; in the majority of cases, however, the default parameters for masking and blurring minimize the extent of the overlaps and thus the potential for spurious high-resolution features in those regions.

3.2.3. Map-based non-uniform prior $m(\mathbf{x})$. Even when no tentative atomic model for the missing structure is available, some rough idea about its placement can be retrieved from the presence of high values of the density (or of its local r.m.s.d.) in noisy electron-density maps, using techniques first developed to perform phase improvement by density modification, either *via* the local average of the electron density (Wang, 1985; Leslie, 1987) or from its local fluctuation around the mean (Reynolds *et al.*, 1985; Jones *et al.*, 1991; Abrahams & Leslie, 1996; Abrahams, 1997).

Once the local density fluctuation, $\omega_\rho(x)$, has been obtained, one may use the homographic exponential model for the whole macromolecular envelope (for details, see Roversi *et al.*, 2000),

$$f_{\rm macrom}(\mathbf{x}) = (1 + \exp\{-\beta_{\rm macrom}[\omega_\rho(\mathbf{x}) - \mu_{\rm macrom}]\})^{-1}. \quad (8)$$

Histogramming of $\omega_\rho(x)$ gives the value of $\mu_{\rm macrom}$ that corresponds to the appropriate solvent fraction, while the value of $\beta_{\rm macrom}$ is taken as proportional to the reciprocal r.m.s. error of the starting density (Blow & Crick, 1959),

$$1/\beta \propto \sum_{\mathbf{h}} \varepsilon_{\mathbf{h}}(1 - {\rm FoM}_{\mathbf{h}}^2)F_{\mathbf{h}}^2, \quad (9)$$

${\rm FoM}_{\mathbf{h}}$ being the figure of merit,

$${\rm FoM}_{\mathbf{h}} = (\langle\cos\varphi_{\mathbf{h}}\rangle^2 + \langle\sin\varphi_{\mathbf{h}}\rangle^2)^{1/2}, \quad (10)$$

computed from the current phase probability distribution $P(\varphi_{\mathbf{h}})$.

Then, to exclude the fragment region from the prior probability distribution for the missing atoms, a homographic exponential model of the fragment density is needed. The local fluctuation, $\omega_\rho^{\rm frag}(\mathbf{x})$, can be computed based on $\rho_{\rm frag}(\mathbf{x})$ as outlined above; the values of $\beta_{\rm frag}$ and $\mu_{\rm frag}$ are computed from the r.m.s. error of the fragment model density and its fractional volume, as seen above. The homographic exponential model for the fragment density is then

$$f_{\rm frag}(\mathbf{x}) = (1 + \exp\{-\beta_{\rm frag}[\omega_\rho^{\rm frag}(\mathbf{x}) - \mu_{\rm frag}]\})^{-1}. \quad (11)$$

Finally, the homographic exponential model for the missing structure envelope is proportional to the probability that position $\mathbf{x}$ lies in the whole macromolecular envelope but not in the fragment envelope,

$$f_{\rm miss}(\mathbf{x}) = f_{\rm macrom}(\mathbf{x})[1 - f_{\rm frag}(\mathbf{x})], \tag{12}$$

and the prior distribution for the placement of the atoms of the missing structure is

$$m^{\rm miss}(\mathbf{x}) = \left[V \Big/ \int_V f_{\rm miss}(\mathbf{x})\,{\rm d}\mathbf{x}\right] f_{\rm miss}(\mathbf{x}). \tag{13}$$

An example of such a map-based missing-atoms distribution, computed for the missing domain 1 of CD55 using a partial atomic model for domains 2, 3 and 4 of CD55, is shown in Fig. 1.

3.2.4. The distribution of $\mathbf{F}_{\mathbf{h}}^{\rm miss}$. Unlike the Luzzati error model, the structure-factor distribution for the missing structure when expanded using an Edgeworth series (Bricogne, 1984) would, strictly speaking, contain terms past the second order. These terms are neglected so that $\mathbf{F}_{\mathbf{h}}^{\rm miss}$ follows a Gaussian distribution around the offset $\mathbf{F}_{\mathbf{h}}^{\rm miss,calc}$ with a variance $\mathbf{V}_{\mathbf{h}}^{\rm miss}$. The structure-factor offset and variance for the missing structure are computed as follows.

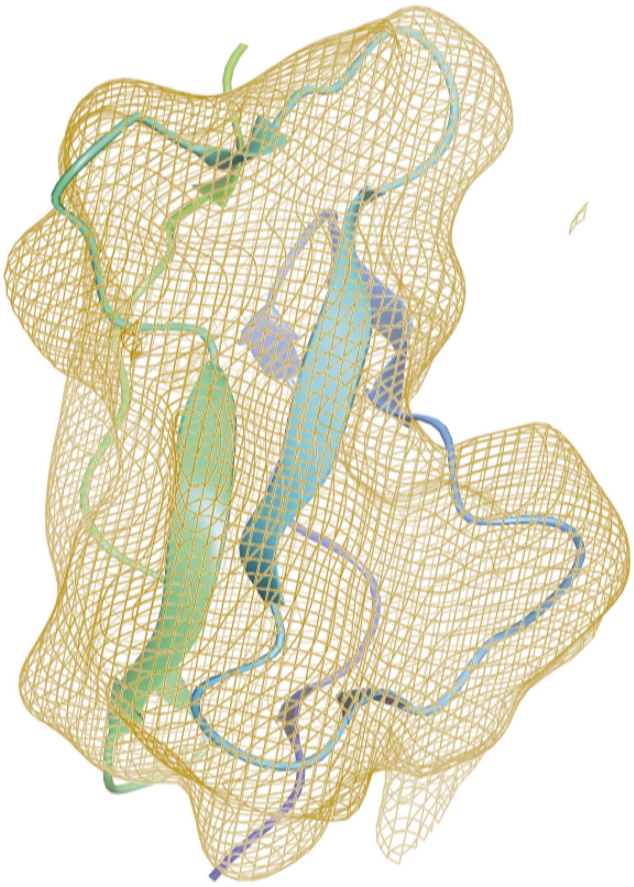

Figure 1
The 3.5σ contour of the *BUSTER* prior distribution, $m^{\rm miss}(\mathbf{x})$, for the missing domain 1 of CD55 in the refinement of $CD55_{234}$ against crystal form *B* $CD55_{1234}$ data, superposed on the model for the same missing domain 1 taken from the deposited structure of the same crystal form of $CD55_{1234}$ (PDB code 1ojw). The distribution is a homographic exponential model (see §3.2.3) computed with variance filtering of a map obtained with the atomic model for CD55 domains 2–4 only (see §7.2.2). The figure was drawn with the program *AESOP* (M. E. M. Noble, personal communication).

(i) The offset is obtained by Fourier transforming the missing structure distribution, $q^{\rm miss}(\mathbf{x})$, and multiplying it by the sum of the missing atoms' scattering factors, $\sigma_1^{\rm miss}(h)$,

$$\mathbf{F}_{\mathbf{h}}^{\rm miss,calc} = \sigma_1^{\rm miss}(h)\,\bar{\mathcal{F}}[q^{\rm miss}(\mathbf{x})](\mathbf{h})\,D_{\mathbf{h}}^{\rm miss}, \tag{14}$$

with $D_{\mathbf{h}}^{\rm miss} = \exp[-(1/4)B_{\rm impf}^{\rm miss}\,d_{\mathbf{h}}^{*2}]$. $\sigma_1^{\rm miss}(h)$ is computed on the basis of an initial guess as to the number, type and *B* factor of the missing atoms,

$$\sigma_1^{\rm miss}(h) = \sum_{j=1}^{N_{\rm miss}} f_j(h)\exp[-(1/4)B_j\,d_{\mathbf{h}}^{*2}]. \tag{15}$$

The ML refinement of a relative scale factor and temperature factor for the missing-atom component (§4.1.2) will correct for errors in this initial guess as to the chemical composition and temperature factors of the missing structure.

(ii) The variance-tensor elements, $\mathbf{V}_{\rm miss}^{ij}$, are computed from the second moment of the missing atoms' unitary structure factor, $\mathbf{U}_{\mathbf{h}}^{\rm miss} = \bar{\mathcal{F}}[q^{\rm miss}(\mathbf{x})](\mathbf{h})$ (using structure-factor algebra; Bertaut, 1955*a*,*b*) and scaled by the imperfection factor and by the inverse square of $K_{\rm impf}^{\rm miss}$,

$$\begin{aligned}
{}^{\mathbf{h}}\mathbf{V}_{\rm miss}^{11} &= \sigma_2^{\rm miss}(h) \times \langle \Re U_{\rm miss}\Re U_{\rm miss}\rangle_{\mathbf{h}} \times [1 - (D_{\mathbf{h}}^{\rm miss})^2]/(K_{\rm impf}^{\rm miss})^2,\\
{}^{\mathbf{h}}\mathbf{V}_{\rm miss}^{12} &= {}^{\mathbf{h}}\mathbf{V}_{\rm miss}^{21}\\
&= \sigma_2^{\rm miss}(h) \times \langle \Re U_{\rm miss}\Im U_{\rm miss}\rangle_{\mathbf{h}} \times [1 - (D_{\mathbf{h}}^{\rm miss})^2]/(K_{\rm impf}^{\rm miss})^2,\\
{}^{\mathbf{h}}\mathbf{V}_{\rm miss}^{22} &= \sigma_2^{\rm miss}(h) \times \langle \Im U_{\rm miss}\Im U_{\rm miss}\rangle_{\mathbf{h}} \times [1 - (D_{\mathbf{h}}^{\rm miss})^2]/(K_{\rm impf}^{\rm miss})^2.
\end{aligned} \tag{16}$$

$\sigma_2^{\rm miss}(h)$ is the sum of missing-structure scattering factors squared,

$$\sigma_2^{\rm miss}(h) = \sum_{j=1}^{N_{\rm miss}} f_j^2(h)\exp[-(1/4)2B_j\,d_{\mathbf{h}}{}^{*2}]. \tag{17}$$

This variance introduces a spatial localization to the 'Wilson' variance traditionally associated with a statistical model for missing atoms; in (16), the variance arising from the statistical nature of the distribution of random atoms is computed using the products between real and imaginary components of the unitary structure factors of the real-space missing-atom distribution.

$\sigma_2^{\rm miss}(h)$ in the same formula represents the average scattering power of the missing atoms, while $K_{\rm impf}^{\rm miss}$ adjusts the 'granularity' of the missing-atom scattering; by the central limit theorem, the missing-atom variance is greater if the scattering comes from a single 'random scatterer' (Bricogne, 1984) of scattering power *f* than if the same amount of scattering is produced by *N* 'random scatterers' of scattering power *f*/*N*, all distributed according to one and the same missing-atom positional probability distribution.

The Luzzati-like variance modulation term $1 - (D_{\mathbf{h}}^{\rm miss})^2$ is introduced in (16) as a means of overcoming the shortcomings arising from neglecting the covariances between the channels. The functional form of this term was selected so as to allow the refinement of the random-atom variance contribution as a function of the resolution. Again, when $B_{\rm impf}^{\rm miss}$ refines towards 0, the variance from the missing structure vanishes, while when

this parameter is large the components of the variance tensor tend toward the full second moments of the missing-atom distribution. Unlike the average intensities that enter the calculation of the fragment (and solvent) variances [see (7) and (20)], the second moments in (16) can extend to high resolution and only the presence of the $\sigma_2^{\rm miss}(\mathbf{h})$ factor ensures the fall-off of the missing structure variance with resolution.

3.3. The bulk solvent

The calculation of the bulk-solvent contribution to the crystal scattering with several different methods of increasing complexity has been described in the crystallographic literature (Glykos & Kokkinidis, 2000, and references therein). The bulk-solvent density in *BUSTER–TNT* is modelled using an envelope uniformly filled with a given solvent electron density, $\bar{\rho}_{\rm s}$, and thermally smeared with a fixed temperature factor $B_{\rm s}$.

3.3.1. The distribution of $\mathbf{F}_{\mathbf{h}}^{\rm solv}$. The structure-factor distribution for the bulk solvent, $\mathbf{F}_{\mathbf{h}}^{\rm solv}$, follows a two-dimensional Gaussian distribution around the offset $\mathbf{F}_{\mathbf{h}}^{\rm solv,calc}$ with a variance $\mathbf{V}_{\mathbf{h}}^{\rm solv}$. The offset and variance are computed as follows.

(i) The offset for the distribution of $\mathbf{F}_{\mathbf{h}}^{\rm solv}$ is the calculated $\mathbf{F}_{\mathbf{h}}^{\rm solv,calc}$,

$$\mathbf{F}_{\mathbf{h}}^{\rm solv,calc} = \bar{\mathcal{F}}[q^{\rm solv}(\mathbf{x})](\mathbf{h}) \times \bar{\rho}_{\rm s} \exp[-(1/4)B_{\rm s}\, d_{\mathbf{h}}^{*2}] \times D_{\mathbf{h}}^{\rm solv}, \quad (18)$$

with $D_{\mathbf{h}}^{\rm solv} = \exp[-(1/4)B_{\rm impf}^{\rm solv}\, d_{\mathbf{h}}^{*2}]$.

$q_{\rm solv}(\mathbf{x})$ is not computed at any stage; the $\bar{\mathcal{F}}[q_{\rm solv}(\mathbf{x})](\mathbf{h})$ term is obtained using the Babinet principle[2] relating the low-resolution Fourier components of two complementary distributions $q_{\rm solv}(\mathbf{x})$ and $q_{\rm macrom}(\mathbf{x})$,

$$\bar{\mathcal{F}}[q^{\rm solv}(\mathbf{x})](\mathbf{h}) = -(V_{\rm macrom}/V_{\rm solv})\bar{\mathcal{F}}[q^{\rm macrom}(\mathbf{x})](\mathbf{h}). \quad (19)$$

$q_{\rm macrom}(\mathbf{x})$ is obtained by masking either around the whole molecule atomic model or around the partial structure[3] and smoothing the resulting binary mask. The masking–smoothing procedure is not detailed here as it is performed in a similar fashion to the procedure described in §3.2.2, where model-based missing structure envelopes are discussed.

$\mathbf{F}_{\mathbf{h}}^{\rm solv,calc}$ should be re-estimated whenever $q_{\rm macrom}(\mathbf{x})$ has changed, typically because the partial structure and/or the missing structure distribution have changed. In the current implementation, however, the bulk-solvent structure factors are computed only once, at the beginning of the *BUSTER* job.[4]

(ii) As seen in the case of the partial structure variance, the solvent-variance tensor, $\mathbf{V}_{\mathbf{h}}^{\rm solv}$, depends on the imperfection parameter $B_{\rm impf}^{\rm solv}$ and is diagonal; its diagonal element, $\mathbf{V}_{\rm solv}^{ii}$, is computed from the average solvent intensity within the resolution bin,

$$^{\mathbf{h}}\mathbf{V}_{\rm solv}^{ii} = (1/2)\,\varepsilon_{\mathbf{h}}\,\langle \bar{\mathcal{F}}[q^{\rm solv}(\mathbf{x})](\mathbf{h})^2 \times \bar{\rho}_{\rm s}^2 \exp[-(1/2)B_{\rm s}\, d_{\mathbf{h}}^{*2}]\rangle_{d^*} \times [1 - (D_{\mathbf{h}}^{\rm solv})^2]. \quad (20)$$

This formulation of the solvent-error model is also chosen by analogy to the Luzzati model for the error in the coordinates of the fragment. This is in spite of the fact that the bulk solvent is not explicitly modelled as atoms, but rather with a real-space electron-density distribution, similar to the missing-atom substructure.

4. The *BUSTER* likelihood function

In the next sections we examine more closely the calculation of the *BUSTER* likelihood function, its gradient and Hessian. First, the dependence of $\mathbf{F}_{\mathbf{h}}^{\rm calc}$ and $\mathbf{V}_{\mathbf{h}}^{\rm calc}$ on the scaling parameters is analysed. We then briefly present the Rice likelihood functions and mention the incorporation of external phases *via* Hendrickson–Lattman (Hendrickson & Lattman, 1970) coefficients.

4.1. Scale factors

All quantities entering a likelihood function need to be on an observational scale; prior to describing the likelihood function, in this section we describe the overall scale and temperature factors that are used to bring quantities from an absolute scale to an observational scale. The values of these overall scale and temperature factors are refined in *BUSTER* by maximizing their likelihood.

The three different contributions to the *BUSTER–TNT* model structure factor may also need to be scaled to one another; relative model scale and temperature factors are also refined jointly with the overall scaling parameters during the ML scaling in *BUSTER*.

4.1.1. The overall scaling and temperature factors. *BUSTER* overall scaling parameters are of three different types: the scale factor $K_{\rm overall}$, the isotropic scaling B factor $B_{\rm iso}$ and the components of the anisotropic scaling tensor β, which enter the isotropic and anisotropic overall scaling factors,

$$F_{\mathbf{h}}^{\rm obs} = [T_{\rm iso}(h)/K_{\rm overall}]T_{\rm aniso}(\mathbf{h})F_{\mathbf{h}}^{\rm obs,abs\ scale}, \quad (21)$$

where (notice the sign of the exponents)

$$T_{\rm iso}(h) = \exp[(1/4)B_{\rm iso}\, d_{\mathbf{h}}^{*2}],$$

$$T_{\rm aniso}(\mathbf{h}) = \exp\left[(1/4)\sum_{ij}\beta_{ij}\mathbf{a}_i^*\cdot\mathbf{a}_j^* h_i h_j\right].$$

The parametrization of the anisotropic scaling factor is slightly unusual but follows the convention adopted in *TNT* (Tronrud *et al.*, 1987; Tronrud, 1997), which constrains the elements of β to make the tensor traceless and, of course, to obey crystal symmetry.

4.1.2. Component-specific scaling parameters. The calculated structure factor is a sum of contributions from three components, individually scaled to one another by individual

[2] For a recent illustration of the use of the Babinet principle in bulk-solvent correction, see Guo *et al.* (2000).
[3] If the partial structure is used instead of the model for the whole macromolecule, the solvent envelope will overlap with the missing structure regions.
[4] This approach makes the *BUSTER–TNT* bulk-solvent correction less adequate after rigid-body refinement, when the atomic shifts are usually large.

scale factors. Because the fragment component usually represents the main contribution to the total structure factor, these scaling factors are all expressed relative to the fragment, which is assumed to be on absolute scale. The expressions for the calculated structure factor, $\mathbf{F}^{\rm calc}$, and its associated variance, $\mathbf{V}_{\mathbf{h}}^{\rm calc}$, on an absolute scale are

$$\mathbf{F}_{\mathbf{h}}^{\rm calc,abs\ scale} = \mathbf{F}_{\mathbf{h}}^{\rm frag,calc} + \frac{\exp[-(1/4)B_{\rm miss}\, d_{\mathbf{h}}^{*2}]}{K_{\rm miss}}\mathbf{F}_{\mathbf{h}}^{\rm miss,calc} + \frac{\exp[-(1/4)B_{\rm solv}\, d_{\mathbf{h}}^{*2}]}{K_{\rm solv}}\mathbf{F}_{\mathbf{h}}^{\rm solv,calc}, \quad (22)$$

$$\mathbf{V}_{\mathbf{h}}^{\rm calc,abs\ scale} = \mathbf{V}_{\mathbf{h}}^{\rm frag,calc} + \frac{\exp[-(2/4)B_{\rm miss}\, d_{\mathbf{h}}^{*2}]}{K_{\rm miss}^{2}}\mathbf{V}_{\mathbf{h}}^{\rm miss,calc} + \frac{\exp[-(2/4)B_{\rm solv}\, d_{\mathbf{h}}^{*2}]}{K_{\rm solv}^{2}}\mathbf{V}_{\mathbf{h}}^{\rm solv,calc}. \quad (23)$$

During scaling, all the component-specific imperfection parameters are refined together with scaling parameters and the full covariance between them is taken into account.

4.2. The Rice distribution and the *BUSTER* likelihood function

The *BUSTER–TNT* distribution of $\mathbf{F}_{\mathbf{h}}$ is a Gaussian centred around the offset $\mathbf{F}_{\mathbf{h}}^{\rm calc}$ (22) with variance $\mathbf{V}_{\mathbf{h}}^{\rm calc}$ (23). When integrated over the phase, it gives the conditional distribution of the structure-factor amplitude, the Rice distribution

$$\mathcal{R}[F_{\mathbf{h}}; \mathbf{F}_{\mathbf{h}}^{\rm calc}(\mathbf{P}), \mathbf{V}_{\mathbf{h}}^{\rm calc}(\mathbf{P})] = \int_{0}^{2\pi} \mathcal{G}[\mathbf{F}_{\mathbf{h}}; \mathbf{F}_{\mathbf{h}}^{\rm calc}(\mathbf{P}), \mathbf{V}_{\mathbf{h}}^{\rm calc}(\mathbf{P})]\, d\varphi_{\mathbf{h}}. \quad (24)$$

The Rice distribution is a function of the set $\mathbf{P}$ of scaling, imperfection and structural parameters, $\mathbf{P} = \{\mathbf{P}_{\rm scal}, \mathbf{P}_{\rm impf}, \mathbf{P}_{\rm struct}\}$, in that these parameters enter the definitions of $\mathbf{F}_{\mathbf{h}}^{\rm calc}$ and $\mathbf{V}_{\mathbf{h}}^{\rm calc}$. The details of the centric and acentric Rice distributions implemented in *BUSTER–TNT* are described by Bricogne (1997).

Assuming independence between reflections, the likelihood of a reflection should be computed by integrating the Rice distribution over the observed structure amplitude, with a probability distribution involving both the observed structure amplitude and its variance. To avoid full two-dimensional integration and to simplify the calculation, for each observed reflection *BUSTER* instead computes the likelihood by consulting the Rice distribution at the value of that observed structure-factor amplitude,

$$\Lambda_{\mathbf{h}}(\mathbf{P}) = \mathcal{R}[F_{\mathbf{h}} = F_{\mathbf{h}}^{\rm obs}; \mathbf{F}_{\mathbf{h}}^{\rm calc}(\mathbf{P}), \mathbf{V}_{\mathbf{h}}^{\rm calc}(\mathbf{P})]. \quad (25)$$

This approach effectively amounts to discarding the uncertainty over the observed structure amplitude. Because the observed uncertainty is usually much smaller than the model error,[5] it is possible to approximate the integration over the observed structure amplitude by adding the observed variance (as a scalar tensor) to the variance obtained from the model.

[5] Only when the model is complete and fairly error free can the magnitudes of model and observation uncertainties be comparable.

The function maximized during the refinement of the parameters is the log-likelihood (LL) of the parameters in view of the observed data, $\mathcal{L}(\mathbf{P})$,

$$\mathcal{L}(\mathbf{P}) = \sum_{\mathbf{h}} \log \Lambda_{\mathbf{h}}(\mathbf{P}). \quad (26)$$

4.3. External phase distribution

Incorporation of external phase information can help refinement, especially in cases of limited resolution and/or data quality (Pannu *et al.*, 1998). When the external phase information is cast in the form of Hendrickson–Lattman *ABCD* coefficients (Hendrickson & Lattman, 1970), its inclusion in the distribution for the structure factor is achieved very simply by adding the 'external' Hendrickson–Lattman coefficients to the 'endogenous' coefficients obtained from the *BUSTER* distribution for the overall structure factor. For this approach to be possible, both phase distributions must share the same origin. The resulting phase probability distribution, $\mathcal{P}_{ABCD}(\varphi)$, is used instead of a uniform weight when integrating over the phase in (24). This process gives rise to the 'elliptic' Rice likelihood function derived by Bricogne (1997).

4.4. The expected structure-factor amplitude $\mathbf{F}_{\mathbf{h}}^{\rm xpct}$

The total structure-factor amplitude can be computed as the first moment of the distribution of the total $\mathbf{F}_{\mathbf{h}}$ and is defined as the expected structure-factor amplitude $F_{\mathbf{h}}^{\rm xpct}$,

$$F_{\mathbf{h}}^{\rm xpct} = \int_{0}^{\infty} \mathcal{R}(F_{\mathbf{h}}) F_{\mathbf{h}}\, {\rm d}F_{\mathbf{h}}. \quad (27)$$

It is $F_{\mathbf{h}}^{\rm xpct}$, on observational scale, that *BUSTER* compares with the observed amplitude $|\mathbf{F}_{\mathbf{h}}|^{\rm obs}$ to compute *R*-factor statistics (see §5.1).

5. Refinement statistics

BUSTER computes several kinds of statistics that serve as a measure of the agreement between the model and the observed data. These statistics are evaluated in resolution bins for free and working sets of reflections and are monitored during the course of the calculation.[6]

5.1. *R* factors

The *R* factors (both overall and in resolution bins) are computed using the expectation of the model structure amplitude rather than its calculated value, on the grounds that the expectation is the current model prediction for an observation. For reflections around resolution d^*,

$$R[d^*] = \langle |F_{\mathbf{h}}^{\rm xpct} - F_{\mathbf{h}}^{\rm obs}| \rangle_{d^*} / \langle |F_{\mathbf{h}}^{\rm obs}| \rangle_{d^*}. \quad (28)$$

[6] Evaluation of the statistics is not performed with a generalized approach (Cowtan, 2002); rather we use conventional binning of reciprocal space in d^2 intervals. Two distinct binnings are effected: a coarser one for overall statistics and another with smaller bin widths for free- and working-sets statistics.

Notice that $F_{\mathbf{h}}^{\rm xpct}$ is on observational scale because it comes from the first moment of the Rice distribution for the structure-factor amplitude (27).

5.2. Log-likelihood gain

To monitor the changes in likelihood of the parameters introduced by the ML refinement, it is useful to consider the log-likelihood gain (Bricogne, 1997), or LLG for short, defined as the logarithm of the ratio between the likelihood of the current parameters over the likelihood of the starting parameters,

$$\mathrm{LLG}(\mathbf{P}_{\rm scal}, \mathbf{P}_{\rm impf}, \mathbf{P}_{\rm struct}) = \log \frac{\Lambda(\mathbf{P}_{\rm scal}, \mathbf{P}_{\rm impf}, \mathbf{P}_{\rm struct})}{\Lambda(\mathbf{P}^0_{\rm scal}, \mathbf{P}^0_{\rm impf}, \mathbf{P}^0_{\rm struct})}. \quad (29)$$

The LLG is zero at the beginning of the calculation and increases whenever the likelihood of the current parameters is higher than the likelihood of the starting parameters. As with R factors, the overall log-likelihood gain is split over the working and the test sets of reflections.

In the context of ML refinement, the log-likelihood gain is a more natural statistic than the R factor and therefore allows a more objective and sensitive assessment of the refinement progress. On the other hand, the LLG cannot inform a comparison of different refinements in that, unlike the familiar R factor, it is relative to the likelihood of the model at the starting cycle.

5.3. Correlation coefficients between structure-factor amplitudes

To check the agreement with the data of structure-factor amplitudes computed from selected subsets of the *BUSTER* structural model, the program computes and outputs correlation coefficients between several kinds of structure-factor amplitudes in resolution bins. Each type of correlation coefficient relates two sets of amplitudes, F_1 and F_2,

$$\mathrm{CC}(F_1, F_2)(d^*) = \frac{\langle F_1F_2\rangle_{d^*} - \langle F_1\rangle_{d^*}\langle F_2\rangle_{d^*}}{[(\langle F_1^2\rangle_{d^*} - \langle F_1\rangle^2_{d^*})(\langle F_2^2\rangle_{d^*} - \langle F_2\rangle^2_{d^*})]^{1/2}}. \quad (30)$$

The correlation coefficients, like the R factors, do not directly contain information as to the quality of the phases, being computed from the amplitudes alone; unlike the R factors, the correlation coefficients are scale-independent. If the refinement and completion are successful, the correlation coefficients should increase (see Fig. 2 for the progression in the case of CD55 refinements, starting from a two-domain only model and ending at the full four-domain structure).

Depending on the particular F_1 and F_2, each correlation coefficient contains information about a specific aspect of data quality, model quality or model-to-data agreement.

(i) $\mathrm{CC}[F_{\rm obs}, F_{\rm obs} + \delta(F_{\rm obs})]$: signal-to-noise ratio. $\delta(F_{\rm obs})$ is a Gaussian random variable of zero mean and s.u. $\sigma(F_{\rm obs})$ and the CC is computed as an expected value over the distribution of these error terms for all reflections. Problems in the data, such as ice rings or drops in the signal-to-noise ratio, are revealed by a sharp decrease in this correlation plot.

(ii) $\mathrm{CC}(F_{\rm obs}, F_{\rm frag})$: how well does the fragment model fit the data? The low-resolution part of the plot shows a dip in the correlation owing to the scattering from the bulk solvent and the missing structure; this scattering contribution is present in $F_{\rm obs}$ but $F_{\rm frag}$ does not account for it. At higher resolution, both incompleteness and fragment imperfection cause loss of correlation. Successful fragment refinement manages to improve this correlation coefficient, especially in the high-resolution range.

(iii) $\mathrm{CC}(F_{\rm obs}, F_{\rm calc})$: how well does the full model fit the data? At low resolution, the inclusion of bulk-solvent and missing-structure models should result in an improved correlation with respect to the $\mathrm{CC}(F_{\rm obs}, F_{\rm frag})$ curve. When refining severely incomplete models, inclusion of the missing-structure model improves the low-resolution correlation (see Fig. 2).

(iv) $\mathrm{CC}(F_{\rm xpct}, F_{\rm calc})$: how adequate is the current error model? When the error model is adequate, the $\mathrm{CC}(F_{\rm xpct}, F_{\rm calc})$ curve loosely follows the $\mathrm{CC}(F_{\rm obs}, F_{\rm calc})$ curve. Overestimated (underestimated) variances lead to a larger (smaller) loss of correlation in $\mathrm{CC}(F_{\rm xpct}, F_{\rm calc})$ than in $\mathrm{CC}(F_{\rm obs}, F_{\rm calc})$.

5.4. $\log\sigma_A$ and overall Luzzati $B_{\rm impf}$

The value of $\log\sigma_A$ (Srinivasan, 1966; Read, 1986) and of the overall Luzzati imperfection, $B_{\rm impf}$, are obtained by computing, respectively, the intercept and slope of the linear regression in d^{*2},

$$(1/2)\log[\langle F^2_{\rm xpct}\rangle/\langle F^2_{\rm obs}\rangle][d^{*2}] = \log\sigma_A + (1/4)B_{\rm impf}d^{*2}. \quad (31)$$

The values of $F_{\rm xpct}$ that enter (31) include contributions from the partial structure, the missing structure and the bulk solvent, so that it is not possible to correlate directly the value of $B_{\rm impf}$ obtained here to the mean-square coordinate error of the partial structure.

6. *BUSTER–TNT* partial structure refinement

Refinement in *BUSTER–TNT* is carried out in much the same way as it is in *TNT* alone, except for three main differences.

(i) The missing-structure contributes a term to the total $\mathbf{F}_{\mathbf{h}}^{\rm xpct}$ and $\mathbf{F}_{\mathbf{h}}^{\rm calc}$.

(ii) The calculated structure factor is handled as a distribution: that is, each $\mathbf{F}_{\mathbf{h}}$ is a vector in the complex plane, distributed around its average value with a model variance that depends on refineable imperfection parameters.

(iii) The scaling, imperfection and structural parameters are refined by maximizing their likelihood in view of the observed structure-factor amplitudes, rather than minimizing the sum of squared differences between observed and calculated structure-factor amplitudes.

BUSTER is the scaling and log-likelihood engine of the calculation, while *TNT* modules perform every other task, called by scripts from within the *BUSTER* binary.

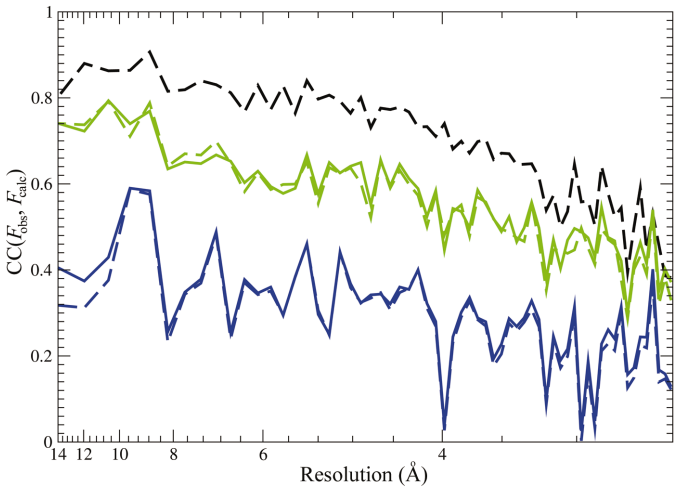

Figure 2
Correlation coefficients, CC(F_{obs}, F_{calc}), between structure-factor amplitudes at the end of *BUSTER–TNT* refinements against the $CD55_{1234}$ data, crystal form *B*, 25–2.8 Å data. Blue curves, refinements of $CD55_{34}$; dashed line, without missing-structure model; full line, with missing-structure model. Green curves, refinements of $CD55_{234}$; dashed line, without missing-structure model; full line, with missing-structure model. Black line, final refinement of the full model for $CD55_{1234}$.

6.1. Algorithmic details

The first task of the *BUSTER* program is to generate the distributions for the missing structure and the solvent.[7] During refinement of a partial structure, those distributions are kept constant, as already mentioned. Overall and component-specific scaling and imperfection parameters and fragment atomic parameters are then refined. At each refinement cycle, the following tasks are performed.

(i) Computation of the fragment structure factors from the current atomic model (in *TNT*).

(ii) Refinement of the scaling and imperfection parameters, while keeping the structural parameters fixed. This task is performed by (i) first assembling the calculated full model structure factors, then (ii) computing the total log-likelihood and its first and second derivatives with respect to all scaling and imperfection parameters and eventually (iii) refining them against ML until convergence (in *BUSTER*).

(iii) The refined scaling and imperfection parameters and the structural model are then used to compute structure-factor amplitude expectation values and all statistics for that cycle (in *BUSTER*).

(iv) The total log-likelihood and its derivatives with respect to the full model structure factors and variances are computed and chained inside *TNT* to generate the gradient and a diagonal approximation to the Hessian matrix of the total log-likelihood with respect to the fragment structural parameters.

(v) These likelihood derivatives are combined with the derivatives of the stereochemical (and optionally NCS) restraints to generate the shift direction for the coordinates, *B* factors and occupancies of the fragment atoms (in *TNT*).

[7] Calls to the *CCP*4 program *NCSMASK* (Collaborative Computational Project, Number 4, 1994) are used to perform any masking steps needed, while any blurring steps are carried out within *BUSTER*

Table 1
Scaling parameters refined by ML in *BUSTER*.

Parameter type	Symbol	Parameters list
Overall scaling	$\mathbf{P}_{scal}$	$K_{overall}$, B_{iso}, β_{ij}
Component scaling	$\mathbf{P}_{scal}$	K_{miss}, B_{miss}, K_{solv}, B_{solv}
Component imperfection	$\mathbf{P}_{impf}$	B_{frag}^{impf}, K_{miss}^{impf}, B_{miss}^{impf}, B_{solv}^{impf}

(vi) Once the step direction has been computed, the step length is optimized by recomputing the total log-likelihood and the restraints residual for models obtained at various step lengths (in *TNT*). Each total log-likelihood evaluation is preceded by an optimization of the scaling and imperfection parameters.

A summary of the parameters involved in scaling is shown in Table 1.

6.2. Approximations in the derivatives

For the sake of computational expediency, some approximations are made while calculating the derivatives of the total log-likelihood during both scaling and structural parameter refinement.

The gradient component of $\mathcal{L}$ for any refined atomic parameter p can be written as

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{\partial \mathcal{L}}{\partial \mathbf{F}^{calc}} \frac{\partial \mathbf{F}^{calc}}{\partial p} + \frac{\partial \mathcal{L}}{\partial \mathbf{V}^{calc}} \frac{\partial \mathbf{V}^{calc}}{\partial p}. \quad (32)$$

BUSTER–TNT neglects the second term in the sum on the right-hand side of (32), that is, the dependence of the variance on the parameter p.[8]

The calculation of the second derivative of the log-likelihood with respect to the partial structure factor needs to accommodate the fact that the *TNT* module *rfactor* can only compute and handle the second derivative of an LS residual, approximately $2/\sigma^2$, while the second derivative of the log-likelihood is needed.

To overcome this limitation, an *ad hoc* $\sigma_{\mathbf{h}}^2$ factor is calculated in *BUSTER* and passed to *TNT*, such that the *TNT* module *rfactor* will effectively compute an approximation to the curvature of the log-likelihood rather than the curvature of the LS residual. The curvature of the log-likelihood can be approximated by a scalar quantity, if we neglect the dependency on the variances and take the average of the absolute values of diagonal elements only,

$$\frac{1}{2}\left[\left|\frac{\partial^2 \mathcal{L}_{\mathbf{h}}}{\partial^2 \Re \mathbf{F}_{\mathbf{h}}^{calc}}\right| + \left|\frac{\partial^2 \mathcal{L}_{\mathbf{h}}}{\partial^2 \Im \mathbf{F}_{\mathbf{h}}^{calc}}\right|\right]\left[\frac{T_{iso}(h) T_{aniso}(\mathbf{h}) D_{\mathbf{h}}^{frag}}{K_{overall}}\right]^2 \quad (33)$$

7. *BUSTER–TNT* refinement of a severely incomplete structure: CD55

As an example of the use of *BUSTER–TNT* to refine and help completion of severely incomplete structures, we will illustrate

[8] The dependence on the variance is not completely neglected, as an entrainment factor is added to $(\partial \mathcal{L}/\partial \mathbf{F}^{calc})\,(\partial \mathbf{F}^{calc}/\partial p)$.

how the program was used to solve the crystal structure of human CD55 starting from a 50% incomplete molecular-replacement model.

CD55 is a four-domain 28 kDa human complement regulator that accelerates the decay of the alternative and classical pathway convertases, thus protecting self-cells from complement-mediated lysis. The crystal structure of a construct consisting of domains 3 and 4 only (hereafter indicated as $CD55_{34}$) was solved first (Williams *et al.*, 2003). Subsequently, crystals of CD55 domains 1–4 (hereafter $CD55_{1234}$) were obtained; they belong to either of two crystal forms: *A* (2.3 Å data; PDB code 1ojv) and *B* (2.8 Å data; PDB code 1ojw). Both forms belong to space group *P*1, with two molecules in the asymmetric unit and about 50% solvent content. For details of data quality and processing, see Lukacik *et al.* (2004).

7.1. Phasing of the $CD55_{1234}$ structure

The molecular-replacement program *MOLREP* (Vagin & Teplyakov, 2000) was used to place two independent copies of the crystallographic model for $CD55_{34}$ (Williams *et al.*, 2003). This model for $CD55_{34}$ was refined in *BUSTER–TNT*, modelling the missing domains 1 and 2 with the homographic exponential model described in §3.2.2.

After refinement and model building on domains 3 and 4, the *BUSTER–TNT* phases were used to locate heavy atoms in a Pt derivative of crystal form *A* and an Au derivative of crystal form *B*. *SHARP* (de La Fortelle & Bricogne, 1997) heavy-atom refinement and phasing of these Pt and Au models

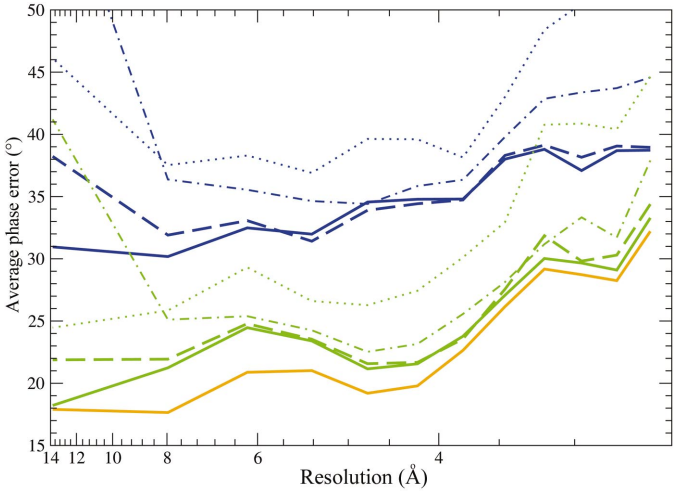

Figure 3
Average phase error $\langle\Delta\varphi\rangle$ in resolution bins for refinements of $CD55_{34}$ and $CD55_{234}$, crystal form *B*, 2.8 Å data. Blue curves, refinements of $CD55_{34}$; dashed line, *BUSTER–TNT* without missing-structure model; full line, *BUSTER–TNT* with missing-structure model; dotted line, *REFMAC5*; dot-dashed line, *CNS*. Green curves, refinements of $CD55_{234}$; dashed line, *BUSTER–TNT* without missing-structure model; full line, *BUSTER–TNT* with missing-structure model; dotted line, *REFMAC5*; dot-dashed line, *CNS*. Orange curve, after maximum-entropy calculation at the end of the *BUSTER–TNT* $CD55_{234}$ refinement that used the missing-structure model.

Table 2
$CD55_{34}$ and $CD55_{234}$ *BUSTER–TNT* overall scale factors.

The ratio to the overall scale factor of the refinement for the full model is reported in parentheses.

Refinement	Crystal form *A*	Crystal form *B*
$CD55_{1234}$	1.137 (1.00)	1.43 (1.00)
$CD55_{34}$, no missing structure	0.82 (0.72)	1.04 (0.73)
$CD55_{34}$	1.24 (1.09)	1.08 (0.76)
$CD55_{234}$, no missing structure	1.00 (0.88)	1.25 (0.87)
$CD55_{234}$	1.07 (0.94)	1.25 (0.87)

in the two crystal forms separately did not lead to interpretable maps (Lukacik *et al.*, 2004).

An iterative phasing procedure was then followed, cycling several times over the following three steps:

(i) phase combination of the *BUSTER–TNT* and *SHARP* phases using *SIGMAA* (Read, 1986);

(ii) NCS and multicrystal averaging across the two crystal forms performed with *DMMULTI* (Cowtan & Main, 1993);

(iii) model building in *XTALVIEW* (McRee, 1999) and new *BUSTER–TNT* refinement.

These steps were repeated until the full structure for domains 1–4 was built and refined in crystal form *A*; the full atomic model for $CD55_{1234}$ built in crystal form *A* was then placed and refined in crystal form *B*.

7.2. Analysis of the $CD55_{34}$ and $CD55_{234}$ *BUSTER–TNT* refinements

In this section, we analyse *BUSTER–TNT* refinements against the $CD55_{1234}$ data. The refinements were performed with and without the missing-structure model at two different stages of model building: the initial refinements of domains 3 and 4 (50% incompleteness) and an intermediate stage where domains 2–4 were built and refined but domain 1 was still missing (25% incompleteness).

7.2.1. $CD55_{34}$. The molecular-replacement solution for the two copies of domains 3 and 4 was rigid-body refined and then subjected to *B*-factor-only refinement, followed by joint positional and *B*-factor refinement with tight NCS restraints in both crystal form *A* and crystal form *B*. After rebuilding of the model for domains 3 and 4, a final round of tight NCS-restrained refinement gave the model for domains 3 and 4 discussed in this section.

At the beginning of each refinement, the low-resolution distribution for the missing domains 1 and 2 was computed as a homographic exponential model (see §3.2.3) based on the nominal solvent content of 50% and variance-filtering of the map obtained from the current phases. The bulk-solvent model was based on the Babinet opposite of the mask around domains 3 and 4 and was therefore overlapping with the missing-structure low-resolution model. In a separate series of refinements, the same protocol was followed in the absence of the missing-structure model, refining the model for domains 3 and 4 with the bulk-solvent model computed, as mentioned above, by masking around domains 3 and 4.

Table 2 reports the overall scale factors for the final refinements of domains 3 and 4, in the presence and absence of the missing-structure model, for both crystal forms of $CD55_{1234}$. Modelling the missing structure clearly helps scaling of the higher-resolution data set, while the improvement for the 2.8 Å data in crystal form *B* is marginal.

However, for both crystal forms, significant improvement in the phases (and in the quality of the derived electron-density and residual maps) is brought about by using the missing-structure model. This improvement is illustrated in Fig. 3 by the plot (for form *B*) of the average phase error,

$$\langle \Delta\varphi \rangle_{d^*}(\{\mathbf{F}_1\}, \{\mathbf{F}_2\}) = \frac{\langle |F_1 F_2(\varphi_1 - \varphi_2)| \rangle_{d^*}}{(\langle F_1^2 F_2^2 \rangle_{d^*})^{1/2}}. \qquad (34)$$

In Fig. 3 we also report the phase error for *REFMAC*5 and *CNS* refinements of the same $CD55_{34}$ model in crystal form *B*; they suffer from a larger phase error, which is expected given the 50% incompleteness and the limited resolution.

7.2.2. $CD55_{234}$. To the refined model for domains 3 and 4, the model for domain 2 was added, taken from a different crystal form (PDB code 1ojy), thus generating a 25% incomplete model for $CD55_{1234}$. Again, *BUSTER–TNT* refinement and rebuilding was carried out separately, with and without a missing-structure model, and the phases and amplitude correlation coefficients were scored at the end. The distribution modelling the missing domain 1 of CD55 based on the phases obtained from the partial atomic model for domains 2, 3 and 4 of CD55, is shown in Fig. 1. The effect of the modelling of the missing structures is still visible in the phase error and phased correlation coefficients plots, whilst the fit to the amplitudes is essentially as good with or without the missing-atoms model (see Fig. 2).

At this level of incompleteness (25%), the phases are good enough to be subjected to a further step of ML phase refinement with maximum entropy constraints in *BUSTER* (Bricogne *et al.*, in preparation), to improve the density for the missing domain 1. The resulting phase error is shown in Fig. 3.

This analysis confirms earlier studies demonstrating that *BUSTER–TNT* can be used successfully to bootstrap refinement and completion from an initial incomplete molecular-replacement solution. At 50% incompleteness, additional phase information is required (in the form of NCS or multi-crystal averaging or poor experimental phases) for structure solution; at 25% incompleteness or lower, the refinement will lead to structure completion, provided the incomplete model is accurate.

Examples of successful use of *BUSTER–TNT* to overcome 10–45% incompleteness and/or reduce phase bias in the refinement of macromolecular models can be found in the literature (*e.g.* Dessen *et al.*, 1999; Fischmann *et al.*, 1999; Bard *et al.*, 2000; Koronakis *et al.*, 2000; Somers *et al.*, 2000; Ng *et al.*, 2000; von Delft *et al.*, 2001; Han *et al.*, 2001; Vicens & Westhof, 2002; Hanzal-Bayer *et al.*, 2002; Benach *et al.*, 2002; Sagermann & Matthews, 2002; Madison *et al.*, 2002; Svensson *et al.*, 2003; Retailleau *et al.*, 2003; Izard *et al.*, 2003).

8. Further developments

Further developments are under way to improve on a number of limitations currently in the software. Among the main improvements planned are

(i) a redesign of the error model, which at the moment suffers from correlations between the bulk-solvent and partial structure errors at low resolution, and suboptimal parameterization of the missing-structure error;

(ii) the use of homographic exponential modelling to compute bulk-solvent envelopes from variance filtering of electron-density maps, to improve the bulk-solvent correction of severely incomplete structures;

(iv) inclusion of off-diagonal terms of the Hessian of the stereochemistry restraints, which will speed up the power of convergence by effectively allowing for joint movement of atoms subjected to bond and angle restraints;

(v) refinement against twinned data;

(vi) PDB deposition tools.

9. Conclusions

BUSTER–TNT offers the possibility of refining incomplete macromolecular atomic models in the presence of a low-resolution probability-based model for the missing structure. The program combines the errors in the partial structure, missing-atoms and bulk-solvent models to give a consistent statistical probability distribution for the structure-factor amplitude, which in turn is used to drive ML refinement of the model.

When the atomic model is very incomplete, modelling of the missing structure and the consistency of the *BUSTER* statistical model help structure building and completion because

(i) the accuracy of the overall scale factors is increased;

(ii) the bias affecting atomic model refinement is reduced by accounting for some of the scattering from the missing structure;

(iii) the addition of a spatial localization to the source of incompleteness improves on traditional Luzzati and sigmaA-based error models;

(iv) the program can perform selective density modification in the regions of unbuilt structure alone.

The program is available for download at http://www.globalphasing.com/buster/.

The authors are grateful to the members of the Global Phasing Consortium for financial support. Partial financial support was also provided by European Commission grant No. QLRT-CT-2000-00398 within the AUTOSTRUCT project. Dale Tronrud helped in the interfacing of *BUSTER* and *TNT*. Eric de La Fortelle, Gwyndaf Evans, Richard J. Morris, Wlodzimierz Paciorek and Marc Schiltz contributed ideas, suggestions and criticism. Very valuable feedback was given by all the beta users of the program and especially by Thierry Fischmann, Sandra Jacob, Dirk Kostrewa and Will Somers. Petra Lukacik expressed, purified and crystallized human

$CD55_{1234}$. The first implementation of the *BUSTER–TNT* scripts was initially written by John J. Irwin. PR is funded by Biotechnology and Biological Sciences Research Council grant No. 43/B16601 (to SML).

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.
Abrahams, J. P. & Leslie, A. (1996). *Acta Cryst.* D**52**, 30–42.
Bard, J., Zhelkovsky, A., Helmling, S., Earnest, T., Moore, C. & Bohm, A. (2000). *Science*, **289**, 1346–1349.
Benach, J., Filling, C., Oppermann, U., Roversi, P., Bricogne, G., Berndt, K., Jornvall, H. & Ladenstein, R. (2002). *Biochemistry*, **41**, 14659–14668.
Bertaut, E. (1955*a*). *Acta Cryst.* **8**, 537–543.
Bertaut, E. (1955*b*). *Acta Cryst.* **8**, 544–548.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* A**12**, 794–802.
Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.
Bricogne, G. (1993*a*). *Acta Cryst.* D**49**, 37–60.
Bricogne, G. (1993*b*). *International Tables for Crystallography*, edited by U. Shmueli, Vol. *B*, pp. 23–106. Dordrecht, Holland: Kluwer Academic Publishers.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 424–448.
Bricogne, G. & Irwin, J. J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (2002). *J. Appl. Cryst.* **35**, 655–663.
Cowtan, K. & Main, P. (1993). *Acta Cryst.* D**49**, 148–157.
Delft, F. von, Lewendon, A., Dhanaraj, V., Blundell, T., Abell, C. & Smith, A. (2001). *Structure*, **9**, 439–450.
Dessen, A., Tang, J., Schmidt, H., Stahl, M., Clark, J., Seehra, J. & Somers, W. (1999). *Cell*, **97**, 349–360.
Fischmann, T., Hruza, A., Niu, X., Fossetta, J., Lunn, C., Dolphin, E., Prongay, A., Reichert, P., Lundell, D., Narula, S. & Weber, P. (1999). *Nature Struct. Biol.* **6**, 233–242.
Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 1070–1072.
Guo, D., Blessing, R. H. & Langs, D. A. (2000). *Acta Cryst.* D**56**, 451–457.
Han, M., Gurevich, V., Vishnivetskiy, S., Sigler, P. & Schubert, C. (2001). *Structure*, **9**, 869–880.
Hanzal-Bayer, M., Renault, L., Roversi, P., Wittinghofer, A. & Hillig, R. (2002). *EMBO J.* **21**, 2095–2106.
Hendrickson, W. & Lattman, E. (1970). *Acta Cryst.* B**26**, 136–143.
Izard, T., Evans, G., Borgon, R., Rush, C., Bricogne, G. & Bois, P. (2003). *Nature* (*London*), **427**, 171–175.
Jones, E., Walker, N. & Stuart, D. (1991). *Acta Cryst.* A**47**, 753–770.
Koronakis, V., Sharff, A., Koronakis, E., Luisi, B. & Hughes, C. (2000). *Nature* (*London*), **405**, 914–919.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 134–136.
Lukacik, P., Roversi, P., White, J., Esser, D., Smith, G., Billington, J., Williams, P., Rudd, P., Wormald, M., Harvey, D., Crispin, M., Radcliffe, C., Dwek, R., Evans, D., Morgan, B., Smith, R. & Lea, S. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1279–1284.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
McRee, D. (1999). *J. Struct. Biol.* **125**, 156–165.
Madison, V. *et al.* (2002). *Biophys. Chem.* **101–102**, 239–247.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Ng, K. K., Petersen, J. F., Cherney, M. M., Garen, C., Zalatoris, J. J., Rao-Naik, C., Dunn, B. M., Martzen, M. R., Peanasky, R. J. & James, M. N. (2000). *Nature Struct. Biol.* **7**, 653–657.
Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Retailleau, P., Huang, X., Yin, Y., Hu, M., Weinreb, V., Vachette, P., Vonrhein, C., Bricogne, G., Roversi, P., Ilyin, V. & Carter, C. J. (2003). *J. Mol. Biol.*, **325**, 39–63.
Reynolds, R., Remington, S., Weaver, L., Fisher, R., Anderson, W., Ammon, H. & Matthews, B. (1985). *Acta Cryst.* B**41**, 139–147.
Roversi, P., Blanc, E., Vonrhein, C., Evans, G. & Bricogne, G. (2000). *Acta Cryst.* D**56**, 1316–1323.
Roversi, P., Irwin, J. J. & Bricogne, G. (1998). *Acta Cryst.* A**54**, 971–996.
Sagermann, M. & Matthews, B. (2002). *J. Mol. Biol.* **316**, 931–940.
Somers, W., Tang, J., Shaw, G. & Camphausen, R. (2000). *Cell*, **103**, 467–479.
Srinivasan, R. (1966). *Acta Cryst.* **20**, 143–144.
Svensson, S., Ostberg, T., Jacobsson, M., Norstrom, C., Stefansson, K., Hallen, D., Johansson, I., Zachrisson, K., Ogg, D. & Jendeberg, L. (2003). *EMBO J.* **22**, 4625–4633.
Tronrud, D. E. (1992). *Acta Cryst.* A**48**, 912–916.
Tronrud, D. E. (1996). *J. Appl. Cryst.* **29**, 100–104.
Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.
Tronrud, D. E. (1999). *Acta Cryst.* A**55**, 700–703.
Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.
Vagin, A. & Teplyakov, A. (2000). *Acta Cryst.* D**56**, 1622–1624.
Vicens, Q. & Westhof, E. (2002). *Chem. Biol.* **9**, 747–755.
Wang, B.-C. (1985). *Methods Enzymol.* **12**, 813–815.
Williams, P., Chaudhry, Y., Goodfellow, I., Billington, J., Powell, R., Spiller, O., Evans, D. & Lea, S. (2003). *Proc. Natl Acad. Sci. USA*, **278**, 10691–10696.