Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Modelling prior distributions of atoms for macromolecular refinement and completion

Pietro Roversi,[a] Eric Blanc,[b] Clemens Vonrhein,[b] Gwyndaf Evans[a] and Gérard Bricogne[a,c]*

[a]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England, [b]Global Phasing Ltd, Sheraton House, Castle Park, Cambridge CB3 0AX, England, and [c]LURE, Université Paris-Sud, Bâtiment 209D, 91405 Orsay, France

Correspondence e-mail: gb10@mrc-lmb.cam.ac.uk

Received 3 May 2000
Accepted 14 June 2000

PDB Reference: porcine pancreatic elastase, 1lvy.

Until modelling is complete, macromolecular structures are refined in the absence of a model for some of the atoms in the crystal. Techniques for defining positional probability distributions of atoms, and using them to model the missing part of a macromolecular crystal structure and the bulk solvent, are described. The starting information may consist of either a tentative structural model for the missing atoms or an electron-density map. During structure completion and refinement, the use of probability distributions enables the retention of low-resolution phase information while avoiding premature commitment to uncertain higher resolution features. Homographic exponential modelling is proposed as a flexible, compact and robust parametrization that proves to be superior to a traditional Fourier expansion in approximating a model protein envelope. The homographic exponential model also has potential applications to *ab initio* phasing of Fourier amplitudes associated with macromolecular envelopes.

1. The case for low-resolution distributions in partial structure refinement and completion

Crystallographic partial structure refinement and completion is usually performed by omitting the questionable parts of the structure and refraining as much as possible from building in ill-defined density regions. If the starting phases are of poor quality, the process of phase improvement by model building is therefore slow, because some of the low-resolution positional information that is already available is not incorporated until the position of the missing atoms is unambiguously defined. In order to avoid locking in on an incorrect structure, even the most likely clues or inspired guesses about the position of the missing atoms are set aside, surrendering to the fear of model bias.

One way of overcoming these difficulties is the iterative placement of atoms in the peaks of the uninterpretable regions of the electron-density map, leading to a 'hybrid model' for the crystal structure that comprises the protein model and free atoms (Perrakis *et al.*, 1999). A different strategy is described here, as implemented in the computer program *BUSTER* (Bricogne, 1993, 1997), which uses a Bayesian statistical model to merge consistently various sources of crystallographic phase information. At any stage during the phasing process, low-resolution real-space distributions are used in *BUSTER* to provide a statistical description of the scattering from the parts of structures that cannot be modelled reliably, either because they are weakly scattering (missing or disordered residues) or because of their intrinsic disorder (bulk solvent).

The main advantages of this procedure are: (i) the scaling of the data to the model is robust and accurate; (ii) the danger of biasing the refinement towards the initial values given to the parameters of the already traced atoms is less serious, because the scattering from the missing atoms is accounted for in a statistical sense; and (iii) from the low-resolution distribution for the missing atoms a maximum-entropy distribution can be derived; suitably scaled and thermally smeared, this is a versatile alternative to conventional weighted difference Fourier maps.

Before we examine closely how the real-space distributions are computed (§4), we add a brief section defining the symbols used throughout (§2) and a section containing the general outline of the structural model as implemented in *BUSTER* (§3).

2. Symbols used in this paper

In this paper, five 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 NX, NY and NZ points along the crystallographic axes. We list here the symbols for these distributions (omitting any subscripts), as an aid to the reader.

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

$\chi(\mathbf{x})$, an indicator function, *i.e.* a binary mask whose values are 0 or 1 only; V_χ is the fractional volume of the mask $\chi(\mathbf{x})$; when the latter is sampled on a crystallographic grid NX NY NZ,

$$V_\chi = (1/\mathrm{NX\,NY\,NZ}) \sum_{i=1}^{NX} \sum_{j=1}^{NY} \sum_{k=1}^{NZ} \chi(i,j,k). \quad (1)$$

$m(\mathbf{x})$, an envelope, *i.e.* a positive everywhere and continuous function, usually with low-resolution Fourier components only; $m(\mathbf{x})$ is normalized so that its average in the unit cell is unity,

$$(1/V) \int_V m(\mathbf{x}) d^3\mathbf{x} = 1, \quad (2)$$

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

$$(1/\mathrm{NX\,NY\,NZ}) \sum_{i=1}^{NX} \sum_{j=1}^{NY} \sum_{k=1}^{NZ} m(i,j,k) = 1. \quad (3)$$

$p(\mathbf{x})$, a probability distribution, so that $0 \leq p(\mathbf{x}) \leq 1$; $\int_V p(\mathbf{x}) d^3\mathbf{x} = 1$.

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

Vertical bars denote the absolute value, $|f(x)| = \mathrm{abs}[f(x)]$; angled brackets denote expectation value under a probability density, $\langle f(\mathbf{x}) \rangle = \int P(\mathbf{x}) f(x)\,\mathrm{d}x$; the asterisk stands for convolution, $(f * g)(x) = \int f(x - y) g(y)\,\mathrm{d}y$.

3. The structural model

The electron density at point $\mathbf{x}$ in the unit cell is written as the sum of three contributions,

$$\rho_{\mathrm{tot}}(\mathbf{x}) = \rho_{\mathrm{frag}}(\mathbf{x}) + \rho_{\mathrm{rand}}(\mathbf{x}) + \rho_{\mathrm{solv}}(\mathbf{x}), \quad (4)$$

where $\rho_{\mathrm{frag}}(\mathbf{x})$ is the electron density for the known *fragment* of the structure for which the atomic positions are known with a good degree of confidence; $\rho_{\mathrm{rand}}(\mathbf{x})$ is the density for the atoms that are missing in the fragment and whose positions are described using a probability distribution and a *random* atom model (see §3.2); $\rho_{\mathrm{solv}}(\mathbf{x})$ is the bulk *solvent* density. Here, $\rho_{\mathrm{tot}}(\mathbf{x})$ is on an absolute scale.

The model for the structure factor is clearly

$$F_{\mathrm{tot}}(\mathbf{h}) = \mathcal{F}[\rho_{\mathrm{tot}}(\mathbf{x})](\mathbf{h}) = F_{\mathrm{frag}}(\mathbf{h}) + F_{\mathrm{rand}}(\mathbf{h}) + F_{\mathrm{solv}}(\mathbf{h}), \quad (5)$$

where the subscripts retain the meaning they have in (4).

Before we describe how the real-space distributions are computed, the next three sections will say some more about the individual components of the structural model.

3.1. The partial structure model

The atoms whose positions are known with a good degree of confidence are described by a set of conventional atomic model parameters. Their positions, isotropic displacement parameters (*i.e.* temperature factors) and occupancies can be refined by maximum likelihood, using an interface to the refinement package *TNT* (Tronrud *et al.*, 1987; Tronrud, 1997), as previously described (Bricogne & Irwin, 1996). The standard stereochemical, geometrical and non-crystallographic symmetry (hard and soft) restraints are handled in *TNT*. During partial structure refinement the probability distribution for the random atoms, as well as the bulk-solvent distribution, are kept fixed.

3.2. The missing structure model

The prior expectation about the position of the missing atoms is cast in quantitative terms using an envelope $m_{\mathrm{rand}}(\mathbf{x})$ that is used as a positional prior distribution for the same atoms; the calculation of $m_{\mathrm{rand}}(\mathbf{x})$ is described in §4. As the suffix 'rand' suggests, all the missing atoms are assumed to be randomly distributed according to $m_{\mathrm{rand}}(\mathbf{x})$.

Once the partial structure has been refined, a maximum-entropy distribution $q_{\mathrm{rand}}(\mathbf{x})$ for the missing atoms is computed in the form

$$q_{\mathrm{rand}}(\mathbf{x}) = \frac{1}{Z} m_{\mathrm{rand}}(\mathbf{x}) \exp\left[\sum_{\mathbf{h}} \lambda_{\mathbf{h}} \Xi_{\mathbf{h}}(\mathbf{x})\right], \quad (6)$$

where Z is a normalization factor such that $\int_V q_{\mathrm{rand}}(\mathbf{x})\,\mathrm{d}^3\mathbf{x} = 1$, $\lambda_{\mathbf{h}}$ are Lagrange multipliers and $\Xi_{\mathbf{h}}$ is the trigonometric structure factor, *i.e.* the structure factor for a point scatterer at rest,

$$\Xi_{\mathbf{h}}(\mathbf{x}) = \frac{1}{|G|} \sum_{g \in G} \exp[2\pi i \mathbf{h} S_g \mathbf{x}]. \quad (7)$$

$|G|$ is the number of elements of the space group G and $S_g\mathbf{x} = \mathbf{R}_g\mathbf{x} + \mathbf{t}_g$ is the generic symmetry operation in G.

The calculation of $q_{\mathrm{rand}}(\mathbf{x})$ is performed varying the $\lambda_{\mathbf{h}}$ under the constraint of maximum entropy, as outlined in Roversi *et al.* (2000).

$q_{\rm rand}(\mathbf{x})$ can be normalized and turned into a positional posterior probability distribution. It shows the extent to which the prior expectation $m_{\rm rand}(x)$ is confirmed or contradicted by the observations. In the absence of noise and if the observations contained no information regarding the region of interest, the final probability distribution would coincide with the (normalized) prior $(1/Z)m_{\rm rand}(\mathbf{x})$ (because $\lambda_{\mathbf{h}} = 0\ \forall\ \mathbf{h}$). In practice, both noise and signal in the data will cause the $\lambda_{\mathbf{h}}$ to differ from zero and build features into $q_{\rm rand}(\mathbf{x})$. The structure-factor contribution to the structure factor from the missing atoms is computed from $q_{\rm rand}(\mathbf{x})$ using the sum of the scattering factors for the same atoms,

$$\mathbf{F}_{\rm rand}(\mathbf{h}) = \Sigma_{\rm rand}(\mathbf{h}) \times \mathcal{F}[\mathbf{q}_{\rm rand}(\mathbf{x})](\mathbf{h}), \tag{8}$$

where $\Sigma_{\rm rand}(\mathbf{h})$ is the sum of the scattering factors for the missing atoms,

$$\Sigma_{\rm rand}(\mathbf{h}) = \sum_{j}^{N_{\rm rand}} f_j(\mathbf{h}) \exp\left[-\langle B\rangle_j \frac{d_{\mathbf{h}}^{*2}}{4}\right]. \tag{9}$$

3.3. The bulk-solvent model

The bulk-solvent structure factor $\mathbf{F}_{\rm solv}(\mathbf{h})$ on the absolute scale can be computed from the Fourier components of the bulk-solvent density $\rho_{\rm solv}(\mathbf{h})$, smeared by the solvent temperature factor,

$$\mathbf{F}_{\rm solv}(\mathbf{h}) = \mathcal{F}[\rho_{\rm solv}(\mathbf{x})](\mathbf{h}) \times \exp\left[-B_{\rm solv} \frac{d_{\mathbf{h}}^{*2}}{4}\right]. \tag{10}$$

The bulk-solvent density is taken proportional to the bulk-solvent envelope $m_{\rm solv}(\mathbf{x})$,

$$\rho_{\rm solv}(\mathbf{x}) = \overline{\rho}_{\rm solv} \times m_{\rm solv}(\mathbf{x}), \tag{11}$$

where $\overline{\rho}_{\rm solv}$ and $V_{\rm solv}$ are the electron density and volume of the bulk solvent.

In *BUSTER*, the bulk-solvent envelope $m_{\rm solv}$ $(\mathbf{x})$ is never handled as such, the macromolecular envelope $m_{\rm macrom}(\mathbf{x})$ being used instead; $m_{\rm macrom}(\mathbf{x})$ is either computed from the whole molecule atomic model [see §4.2, the volume $V_{\rm macrom}(\mathbf{x})$ being the volume of the whole binary mask $\chi_{\rm macrom}(\mathbf{x})$] or it is computed starting from the density using the known solvent-volume fraction (see §4.3).

Once $m_{\rm macrom}(\mathbf{x})$ is obtained, the Babinet principle,[1] relating the low-resolution Fourier components of two complementary distributions $m_{\rm solv}(\mathbf{x})$ and $m_{\rm macrom}(\mathbf{x})$, is used,

$$V_{\rm solv}\mathcal{F}[m_{\rm solv}(\mathbf{x})](\mathbf{h}) = -V_{\rm macrom}\mathcal{F}[m_{\rm macrom}(\mathbf{x})](\mathbf{h}), \tag{12}$$

so that

$$\mathbf{F}_{\rm solv}(\mathbf{h}) = -\overline{\rho}_{\rm solv}V_{\rm macrom} \times \mathcal{F}[m_{\rm macrom}(\mathbf{x})](\mathbf{h}) \times \exp\left[\frac{(-d_{\mathbf{h}}^{*})^2}{4} B_{\rm solv}\right]. \tag{13}$$

[1] For a recent illustration of the use of the Babinet principle in bulk-solvent correction, see Guo *et al.* (2000).

4. Computing $m_{\rm rand}(\mathbf{x})$

We can now examine more closely how the real-space envelopes are computed; in particular, we discuss here the calculation of the envelope for the missing atoms, $m_{\rm rand}(\mathbf{x})$. Similar techniques can be used to compute the envelopes for the whole macromolecule or for the bulk solvent.

As soon as an initial model is available, the prior distribution $m_{\rm rand}(\mathbf{x})$ for the positions of the missing atoms can be computed in three ways: (i) by excluding the missing atoms from the regions already containing the partial structure (uniform prior, §4.1), (ii) by using a trial atomic model for the missing atoms (model-based non-uniform prior, §4.2) or (iii) simply from the local fluctuation of the electron density (map-based non-uniform prior, §4.3).

4.1. Uniform prior

The simplest choice for the missing atoms prior probability distribution is to exclude them from the regions that already contain a reliable atomic model: this 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 three steps as follows.

(i) A binary mask $\chi_{\rm frag}^{\rm a.u.}(\mathbf{x})$ is drawn around the known partial structure; this step is performed using the program *NCSMASK* (Collaborative Computational Project, Number 4, 1994). The masking radius $R_{\rm frag}$ can be varied; the default for $R_{\rm frag}$ is 2.05 Å.

(ii) $\chi_{\rm frag}^{\rm a.u.}(\mathbf{x})$ is symmetry expanded to cover the whole cell; this symmetry-expanded binary mask $\chi_{\rm frag}(\mathbf{x})$ is negated to obtain a binary mask $\chi_{\rm rand}(\mathbf{x})$ for the random atoms,

$$\chi_{\rm rand}(\mathbf{x}) = 1 - \chi_{\rm frag}(\mathbf{x}). \tag{14}$$

(iii) $\chi_{\rm frag}(\mathbf{x})$ is blurred by means of a convolution with an isotropic Gaussian $G(\mathbf{x}; B_{\rm rand})$ and normalized,

$$m_{\rm rand}(\mathbf{x}) = \frac{1}{V_{\chi_{\rm rand}}} \times [\chi_{\rm rand} * G(B_{\rm rand})](\mathbf{x}), \tag{15}$$

where the parameter $B_{\rm rand}$ controls the width of the Gaussian and therefore the slope of $m_{\rm rand}(\mathbf{x})$ around the model used in generating $\chi_{\rm frag}^{\rm a.u.}(\mathbf{x})$.

The convolution in (15) is effected in reciprocal space, using a set of periodized ('aliased') structure factors for $m_{\rm rand}(\mathbf{x})$. The use of aliased structure factors to sample thermally smeared model densities on arbitrarily coarse crystallographic grids has been described in the Appendix of Roversi *et al.* (1998) and will not be detailed here.[2]

[2] Suffice here to say that first $\mathcal{F}[m_{\rm rand}(\mathbf{x})](\mathbf{h})$ is computed by taking the products of $\mathcal{F}[\chi_{\rm rand}(\mathbf{x})](\mathbf{h})$ and $\mathcal{F}[G(\mathbf{x};\ B_{\rm frag})](\mathbf{h})$; then, the set of $\mathcal{F}[m(\mathbf{x})_{\rm rand}](\mathbf{h})$ are made periodic on the lattice reciprocal to the real-space crystallographic grid. These aliased structure factors undergo Fourier synthesis and $m_{\rm rand}(\mathbf{x})$ is sampled on the desired grid; the aliasing ensures that the $m_{\rm rand}(\mathbf{x})$ distribution is positive everywhere 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 absence of any partial model is this a truly uniform distribution throughout the unit cell.

We also notice that if the bulk-solvent envelope is also chosen to fill up all the space left empty by the macromolecular model, the missing atoms envelope and the bulk-solvent envelope are overlapping. They can still differ for the parameter B used in the blurring step (15).

4.2. Model-based non-uniform prior

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 $m_{\rm rand}(\mathbf{x})$ can then be built around these ill-defined atoms and the same atoms 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 atoms 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 four steps as follows.

(i) A binary mask $\chi^{\rm a.u.}_{\rm macrom}(\mathbf{x})$ is drawn around the complete atomic model, including the parts that will be omitted; the radius for this masking can vary between 2 and 4 Å, depending on the degree of confidence one wants to retain regarding the omitted model (a tighter radius resulting in a distribution highly localized around the omitted atoms).

(ii) A binary mask $\chi^{\rm a.u.}_{\rm frag}(\mathbf{x})$ is drawn around the part of structure that is going to be retained and a binary mask for the random atoms $\chi^{\rm a.u.}_{\rm rand}(\mathbf{x})$ is obtained from

$$\chi^{\rm a.u.}_{\rm rand}(\mathbf{x}) = \chi^{\rm a.u.}_{\rm macrom}(\mathbf{x}) \times \left[1 - \chi^{\rm a.u.}_{\rm frag}(\mathbf{x})\right]. \qquad (16)$$

(iii) The $\chi^{\rm a.u.}_{\rm rand}(\mathbf{x})$ mask is symmetry expanded to the unit cell to give $\chi_{\rm rand}(\mathbf{x})$.

(iv) $\chi_{\rm rand}(\mathbf{x})$ is blurred by means of a convolution with an isotropic Gaussian $G(\mathbf{x}; B_{\rm rand})$ and normalized as in (15).

4.3. Map-based non-uniform prior

Even when no atomic model is available, some rough idea about the placement of the missing atoms can be retrieved from the presence of high values of the local r.m.s.d. in noisy electron-density maps.

The local average of the electron density (Wang, 1985; Leslie, 1987) or its local fluctuation around the mean (Abrahams & Leslie, 1996; Abrahams, 1997) have been used to perform phase improvement by density-modification techniques.

The *BUSTER* envelope is also computed by local variance filtering of a noisy density map. Local averaging is performed by convolution with a Gaussian $G(B)$, parametrized by a Debye–Waller factor B, and a solid sphere mask $S(R)$, parametrized by a radius R. These convolutions are used in two filtering operations that select high and low frequencies in a distribution $\rho(\mathbf{x})$,

$$\rho^{\rm lo}(B, R)(\mathbf{x}) = [\rho * G(B) * S(R)](\mathbf{x}) \qquad (17)$$

$$\rho^{\rm hi}(B, R)(\mathbf{x}) = (\rho - \rho^{\rm lo})(\mathbf{x}). \qquad (18)$$

All the convolution steps are carried out in reciprocal space, by calculation of a set of aliased structure factors (Roversi *et al.*, 1998), then Fourier-transformed to sample the density on the required grid.

For the (optional) high-frequency filtering, the following two measures of the local fluctuation around the local average can be defined:

(i) the local average of the absolute value of the deviation from the mean,

$$\omega(\mathbf{x}) = [|\rho^{\rm hi}(B_1, R_1)| * G(B_2) * S(R_2)](\mathbf{x}), \qquad (19)$$

(ii) the local r.m.s.d. from the local average,

$$\omega(\mathbf{x}) = \{[\rho^{\rm hi}(B_1, R_1)]^2 * G(B_2) * S(R_2)\}^{\frac{1}{2}}(\mathbf{x}). \qquad (20)$$

The radius of the sphere for the high-pass filter is typically larger than the one for the low-pass filter in (19) and (20) (*i.e.* $R_1 > R_2$).

The high-frequency filter is useful in those cases where map Fourier components with $D \leq R_1$ are either absent or cannot be trusted; but it can be omitted if the lowest-resolution features are correct; in this case, the following two local averages can be computed, also by Fourier transforms:

(i) the local average of the absolute value of the density,

$$\omega(\mathbf{x}) = [|\rho^{\rm lo}| * G(B_2) * S(R_2)](\mathbf{x}), \qquad (21)$$

(ii) the local r.m.s. deviation from zero of the density,

$$\omega(\mathbf{x}) = [(\rho^{\rm lo})^2 * G(B_2) * S(R_2)]^{\frac{1}{2}}(\mathbf{x}). \qquad (22)$$

Once $\omega(\mathbf{x})$ is available, $m_{\rm rand}(\mathbf{x})$ should be obtained by homographic exponential modelling as described in the following section.

5. Homographic exponential modelling

We describe in this section a technique that affords a parametrization of low-resolution distributions and is used in *BUSTER* for computing macromolecular envelopes from noisy electron-density maps. The technique is a particular case of *homographic* mapping of a function $e(\mathbf{x})$,

$$e(\mathbf{x}) \rightarrow \frac{a + b \times e(\mathbf{x})}{c + d \times e(\mathbf{x})}, \qquad (23)$$

where $a = c = d = 1$ and $b = 0$, and $e(\mathbf{x})$ is an *exponential* $e(\mathbf{x}) = \exp[\omega(\mathbf{x})]$; therefore, we propose to call it *homographic exponential modelling*.

The distributions obtained by homographic exponential modelling can be handled as values on a crystallographic grid and represent a new way of defining intrinsically 'binary-like' macromolecular envelopes that are continuous and not binary. Alternatively, they can be parametrized with a finite set of

coefficients in the expansion of ω, opening the way to *ab initio* low-resolution phasing based on phase permutation for a few coefficients of $\omega(\mathbf{x})$.

The potential of the homographic exponential modelling for *ab initio* phasing of envelope Fourier coefficients has been investigated by G. Bricogne and M. Ramin (G. Bricogne, unpublished results; Ramin, 1999). Here, we introduce the technique and present the results of a test study, aiming at the assessment of the number of Fourier coefficients of $\omega(\mathbf{x})$ that are needed to satisfactorily reconstruct a given $m(\mathbf{x})$ when a homographic exponential model is adopted.

5.1. The Fermi–Dirac distribution

The problem of defining a low-resolution envelope for the macromolecule based on an electron-density map can be restated in the form of assigning to each pixel in the map a probability of belonging to the bulk solvent, which we can write $p_{\rm solv}(\mathbf{x})$. Correspondingly, $p_{\rm macrom}(\mathbf{x}) = 1 - p_{\rm solv}(\mathbf{x})$ is then the probability that the pixel at $\mathbf{x}$ belongs to the macromolecular volume.

It is clear that we are dealing with each pixel as an entity that can be in one and one only of two possible states (pixel in the bulk solvent/pixel in the macromolecule), like a fermion whose spin can be either of $\pm\frac{1}{2}$; an analogy can be drawn with the occupancy distribution function for a system consisting of a finite number of fermion particles with a given total energy. This occupancy distribution function $f_{\rm FD}(E)$ follows a Fermi–Dirac distribution, depending on the temperature parameter $\beta_{\rm FD}$ and on the chemical potential $\mu_{\rm FD}$ (Reif, 1965),

$$f_{\rm FD}(E) = 1/\{1 + \exp[\beta_{\rm FD}(E - \mu_{\rm FD})]\}. \qquad (24)$$

The chemical potential $\mu_{\rm FD}$ arises from the requirement that the number of fermions is finite. At temperatures close to zero, the low-energy states are occupied [probability $f_{\rm FD}(E) \simeq 1$] until the total number of fermions is reached; this defines the Fermi level (or Fermi energy $\mu_{\rm FD}$) of the system. The distribution quickly tails off to zero as the energy level increases; the states having energy higher than the Fermi level have zero occupancies unless the ratio of the energy gap $(E - \mu_{\rm FD})$ over the mean thermal energy $1/\beta_{\rm FD}$ is small enough to permit some excitation.

By analogy, we can adopt some measure $\omega(\mathbf{x})$ of the local fluctuation of the electron density as an 'envelope potential energy' and take β as inversely proportional to the r.m.s. error of the electron density (Blow & Crick, 1959),

$$\frac{1}{\beta} \propto \sum_{\mathbf{h}} \varepsilon_{\mathbf{h}}\left(1 - {\rm FOM}_{\mathbf{h}}^2\right)F_{\mathbf{h}}^2, \qquad (25)$$

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

$${\rm FOM}_{\mathbf{h}} = (\langle\cos\varphi_{\mathbf{h}}\rangle^2 + \langle\sin\varphi_{\mathbf{h}}\rangle^2)^{1/2}, \qquad (26)$$

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

Where $\omega(\mathbf{x})$ is large with respect to the density r.m.s. error, it is highly unlikely that pixel $\mathbf{x}$ belong to the bulk solvent. So, for the probability that the pixel belong to the solvent, we can take

$$p_{\rm solv}(\mathbf{x}) \propto \frac{1}{1 + \exp\{\beta[\omega(\mathbf{x}) - \mu]\}}. \qquad (27)$$

The value of μ depends on the number of pixels that define the solvent region (or the solvent-volume fraction); it can be computed by histogramming the $\omega(\mathbf{x})$ function and choosing for μ the value of $\omega(\mathbf{x})$ that will give the correct number of pixels within the solvent, starting from the pixels where the fluctuation is lowest, and including all the pixels with increasing values of the local fluctuation, until the desired solvent fraction is achieved.

The probability that the pixel at $\mathbf{x}$ belongs to the macromolecule is then

$$p_{\rm macrom}(\mathbf{x}) = 1 - p_{\rm solv}(\mathbf{x}) \propto \frac{1}{1 + \exp\{-\beta[\omega(\mathbf{x}) - \mu]\}}. \qquad (28)$$

5.2. Homographic exponential modelling of missing atoms envelopes

This section describes the homographic exponential modelling of macromolecular envelopes starting from noisy maps. In particular, a description is given of the calculation of an homographic exponential model for the missing atom envelope in the presence of the density for the partial structure $\rho_{\rm frag}(\mathbf{x})$ (see §4.3).

Once the local density fluctuation $\omega(\mathbf{x})$ has been obtained along the lines described in §4.3 and its histogramming has given the value of $\mu_{\rm macrom}$ that corresponds to the appropriate

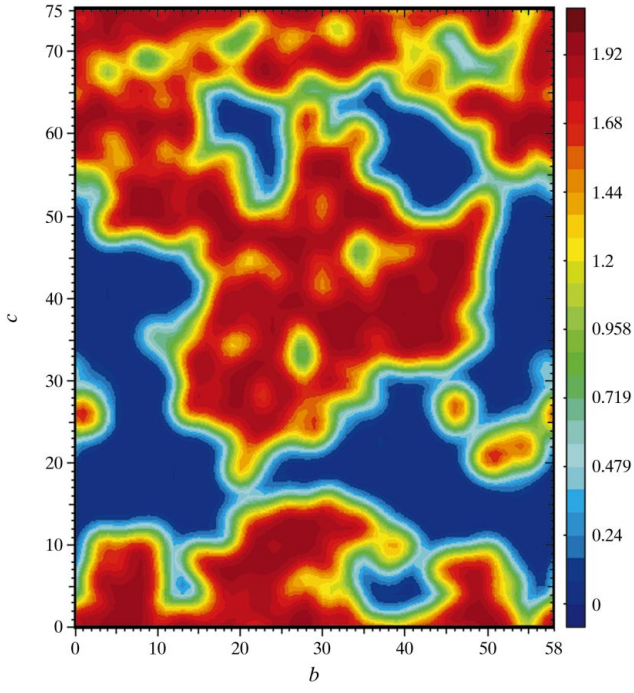

Figure 1
Porcine pancreatic elastase, [100] section of the model envelope $m(\mathbf{x})$. Section: 57.973 × 75.32 Å. The centre of the section is the macromolecule's centre of gravity. The density was obtained by masking with a radius of 2 Å around the model and blurring with a Gaussian temperature factor B = 100.

solvent fraction, one has the homographic exponential model for the whole macromolecular envelope,

$$q_{\rm macrom}(\mathbf{x}) = \frac{1}{1+\exp\{-\beta_{\rm macrom}[\omega(\mathbf{x})-\mu_{\rm macrom}]\}}, \qquad (29)$$

the value of $\beta_{\rm macrom}$ being proportional to the reciprocal r.m.s. error of the starting density (25). Then, to exclude the fragment region from the prior-probability distribution for the random atoms, a homographic exponential model of the fragment density is needed. The local fluctuation $\omega_{\rm frag}(\mathbf{x})$ can be computed based on $\rho_{\rm frag}(\mathbf{x})$ as outlined in §4.3; 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

$$q_{\rm frag}(\mathbf{x}) = \frac{1}{1+\exp\{-\beta_{\rm frag}[\omega_{\rm frag}(\mathbf{x})-\mu_{\rm frag}]\}}. \qquad (30)$$

Finally, the homographic exponential model for the missing atoms envelope is obtained by imposing that the pixel lies in the whole macromolecule envelope but not in the fragment envelope,

$$q_{\rm rand}(\mathbf{x}) = q_{\rm macrom}(\mathbf{x}) \times \left[1 - q_{\rm frag}(\mathbf{x})\right] \qquad (31)$$

$$m_{\rm rand}(\mathbf{x}) = \frac{V}{\int_V q_{\rm rand}(\mathbf{x})\,{\rm d}^3\mathbf{x}} \times q_{\rm rand}(\mathbf{x}). \qquad (32)$$

5.3. A simple test

We describe here a simple calculation that investigates the behaviour of homographic exponential modelling of a known envelope $m(\mathbf{x})$ under truncation of its Fourier spectrum, and compares it with a traditional finite-resolution Fourier expansion of the same $m(\mathbf{x})$.

If $m(\mathbf{x})$ is a given envelope and we intend to parametrize it using an homographic exponential model (28), we first map $m(\mathbf{x})$ to the (0, 1) open interval by linear scaling,

$$m'(\mathbf{x}) = \frac{[m(\mathbf{x}) - \min m(\mathbf{x})]}{[\max m(\mathbf{x}) - \min m(\mathbf{x})]}. \qquad (33)$$

Then, we can compute the $\omega(\mathbf{x})$ from

$$\omega(\mathbf{x}) = \frac{1}{\beta}\log\left[\frac{m'(\mathbf{x})}{1-m'(\mathbf{x})}\right] + \mu. \qquad (34)$$

Fourier analysis of $\omega(\mathbf{x})$, truncation of its Fourier coefficients at resolution d and Fourier synthesis of the truncated set of coefficients lead to the resolution-truncated $\omega_d(\mathbf{x})$ distribution

$$\omega_d(\mathbf{x}) = \overline{\mathcal{F}}\{X_d(\mathbf{h}) \times \mathcal{F}[\omega(\mathbf{x})](\mathbf{h})\}(\mathbf{x}), \qquad (35)$$

where the truncation of the Fourier spectrum of $\omega(\mathbf{x})$ at resolution d in (35) is performed by multiplying it by the indicator function $X_d(\mathbf{h})$,

$$\begin{aligned} X_d(\mathbf{h}) &= 1 \text{ if } h \geq d, \\ &= 0 \text{ if } h < d. \end{aligned} \qquad (36)$$

The homographic exponential, resolution-truncated $m_{{\rm HE},d}(\mathbf{x})$ is then

Table 1
Porcine pancreatic elastase: real-space correlation coefficients between a model envelope $m(\mathbf{x})$ and its reconstructions by truncated homographic exponential modelling $[m_{{\rm HE},d}(\mathbf{x})]$ and truncated Fourier synthesis $[m_{{\rm FT},d}(\mathbf{x})]$.

Resolution d (Å) (No. coeffs)	$\langle {\rm CC}(m, m_{{\rm HE},d})\rangle$	$\langle {\rm CC}(m, m_{{\rm FT},d})\rangle$
30 (7)	0.594	0.604
25 (12)	0.634	0.662
20 (22)	0.760	0.758
15 (51)	0.840	0.832

$$m'_{{\rm HE},d}(\mathbf{x}) = \frac{1}{1+\exp\{-\beta[\omega_d(\mathbf{x})-\mu]\}}, \qquad (37)$$

$$m_{{\rm HE},d}(\mathbf{x}) = \frac{V}{\int_V m'_{{\rm HE},d}(\mathbf{x})\,{\rm d}^3\mathbf{x}} \times m'_{{\rm HE},d}(\mathbf{x}). \qquad (38)$$

We note here that for this particular test the actual values of β and μ are irrelevant, provided the same values are used in (34) and (37).

The conventional Fourier expansion of $m(\mathbf{x})$, with truncation at resolution d, reads

$$m_{{\rm FT},d}(\mathbf{x}) = \overline{\mathcal{F}}\{X_d(\mathbf{h}) \times \mathcal{F}[m(\mathbf{x})](\mathbf{h})\}(\mathbf{x}). \qquad (39)$$

$m_{{\rm HE},d}(\mathbf{x})$ and $m_{{\rm FT},d}(\mathbf{x})$ differ from $m(\mathbf{x})$ because of the resolution truncation; $m_{{\rm FT},d}(\mathbf{x})$ has no Fourier components past d Å, while $m_{{\rm HE},d}(\mathbf{x})$, computed from the same number of Fourier coefficients, possesses extra-resolution owing to the exponential step.

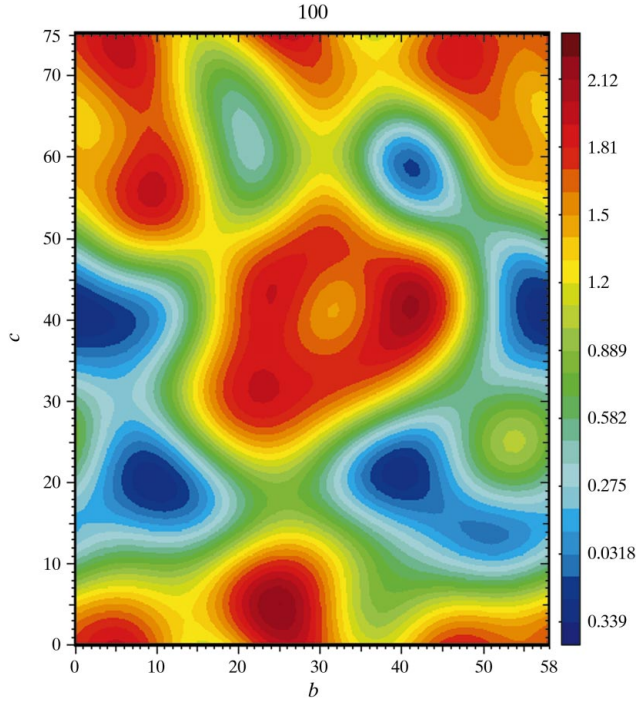

Figure 2
Porcine pancreatic elastase, [100] section of the 15 Å truncated Fourier reconstruction of the model envelope, $m_{{\rm FT},d=15\text{Å}}(\mathbf{x})$. Size and orientation as in Fig. 1. The density was obtained by truncating the Fourier spectrum of the model density at 15 Å [51 data; see (39)].

In the following, we describe the test reconstruction of a model envelope for porcine pancreatic elastase (PPE; Meyer *et al.*, 1986; Schiltz *et al.*, 1997). The model envelope $m(\mathbf{x})$ was generated as explained in §4.2, using the PDB-deposited structure, with a masking radius $R = 2$ Å and a blurring factor $B = 100$. A conventional Fourier truncation and a truncated homographic exponential model were used to reconstruct the model envelope, as explained above. As noted in §2, all envelopes have been normalized so that their average in the unit cell is unity.

Table 1 reports the real-space overall correlation coefficients between the model envelope and its Fourier-truncated and homographic exponential-truncated reconstructions. The Fourier-truncated envelope gives marginally higher CCs when the resolution used for truncating the coefficients is lower than 25 Å: this is because the amplitudes and phases of the very few coefficients retained are exact for this envelope and not for $m_{\mathrm{HE},d}(\mathbf{x})$. Overall, the values of the CCs are very similar for the two methods, mainly because the correlation coefficients are dominated by the lowest resolution components, which are essentially correct in both maps.

More informative is the visual inspection of sections of the envelopes. Fig. 1 shows a section in the [100] plane of the PPE crystal for the model envelope; Figs. 2 and 3 show the same section of the 15 Å, Fourier-truncated and homographic exponential truncated envelopes, respectively, $m_{\mathrm{FT},d=15\text{Å}}(\mathbf{x})$ and $m_{\mathrm{HE},d=15\text{Å}}(\mathbf{x})$. In Fig. 2, $m_{\mathrm{FT},d=15\text{Å}}(\mathbf{x})$ shows the well known Fourier artefacts arising from truncation: negative ripples, peaky features and a smeared out protein–solvent boundary.

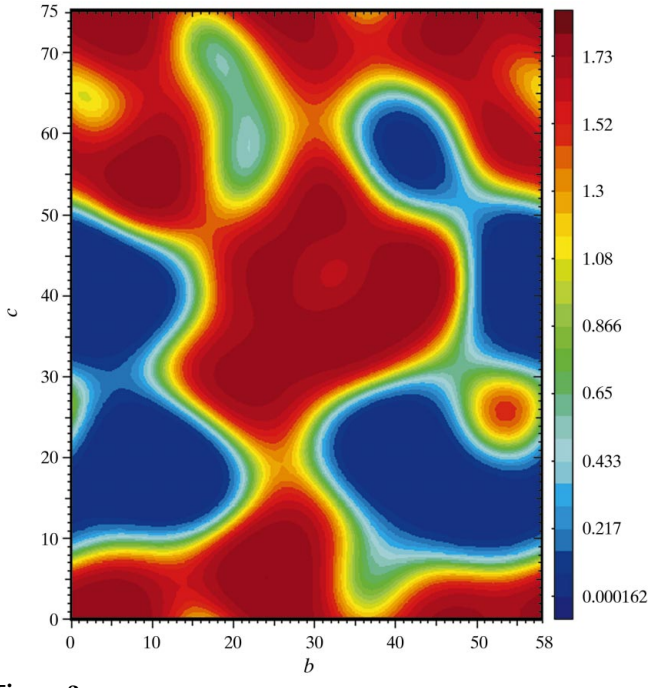

Figure 3
Porcine pancreatic elastase, [100] section of the 15 Å truncated homographic exponential reconstruction of the model envelope, $m_{\mathrm{HE},d=15\text{ Å}}(\mathbf{x})$. Size and orientation as in Fig. 1. The density was obtained by truncating the ω spectrum at 15 Å (51 data) and recomputing the homographic exponential model (37).

Table 2
Porcine pancreatic elastase: reciprocal-space correlation coefficients between the Fourier components $\mathcal{F}[m(\mathbf{x})](\mathbf{h})$ of a model envelope and the Fourier components $\mathcal{F}[_{m\mathrm{HE},d}(\mathbf{x})](\mathbf{h})$ of its truncated homographic exponential reconstruction.

	$\langle \mathrm{CC}\{\mathcal{F}[m(\mathbf{x})](\mathbf{h}), \mathcal{F}[m_{\mathrm{HE},d}(\mathbf{x})](\mathbf{h})\}\rangle$	
Resolution (Å) (No. coeffs)	$d = 15$ Å	$d = 20$ Å
14.1 (61)	0.982	0.920
10.0 (93)	0.170	0.125
8.2 (125)	0.306	0.087
7.1 (136)	0.118	−0.007
6.3 (151)	0.042	−0.040
5.8 (166)	0.154	0.079

In Fig. 3, $m_{\mathrm{HE},d=15\text{Å}}(\mathbf{x})$ is positive everywhere, has a flatter protein ceiling, a steeper slope at the solvent–protein boundary and a flatter solvent floor, with few oscillations. The solvent regions match the ones in the model envelope.

Table 2 contains the correlation coefficients between Fourier coefficients of the model PPE envelope and the Fourier coefficients of the 15 and 20 Å truncated homographic exponential model. Fig. 4 plots the same Fourier coefficients in resolution ranges. The fluctuations observed are typical of the spectrum of macromolecular envelopes; still, the amplitudes of the Fourier components of $m_{\mathrm{HE},d=15\text{Å}}(\mathbf{x})$ retain an average correlation coefficients as high as 0.306 up to 8.2 Å, owing to the extrapolation achieved by the exponential step.

6. Conclusions

The macromolecular envelope $m_{\mathrm{rand}}(\mathbf{x})$ is a *continuous* distribution and not a binary mask; even regions of low density (or low-density r.m.s.d., if a variance filter is used) can therefore be retained within the envelope, with a (possibly small) non-zero probability. The subsequent maximum entropy modulation of the envelope itself therefore has a chance of

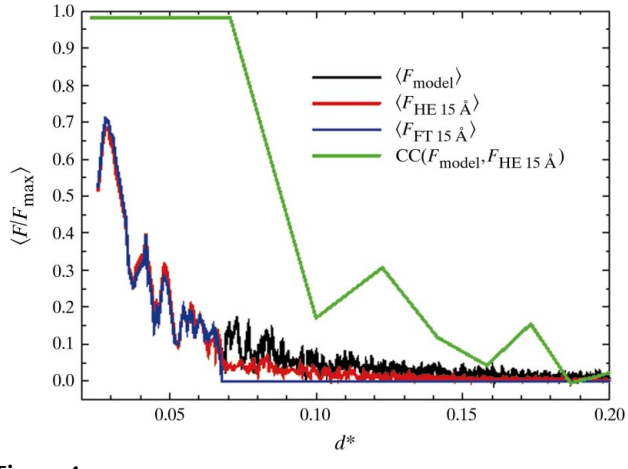

Figure 4
Porcine pancreatic elastase. Fourier components of the model envelope $\langle\mathcal{F}[m(\mathbf{x})](\mathbf{h})\rangle$ and of its 15 Å truncated reconstructions $\langle\mathcal{F}[m_{\mathrm{FT}}(\mathbf{x})](\mathbf{h})\rangle$ and $\langle\mathcal{F}[m_{\mathrm{HE}}(\mathbf{x})](\mathbf{h})\rangle$. *F*s were averaged in groups of ten data each. The correlation coefficients $\langle\mathrm{CC}\{\mathcal{F}[m(\mathbf{x})](\mathbf{h}), \mathcal{F}[m_{\mathrm{FT}}(\mathbf{x})](\mathbf{h})\}\rangle$ are not shown because they are 1.0 for $d > 15$ Å and zero for $d < 15$ Å.

building up density in the same regions. This has potential in structure completion by density-modification techniques. The only other published example of solvent flattening using real-space continuous probability distributions is the Gaussian distribution described by Terwilliger (1999). The map-based algorithm implemented in *BUSTER* (§5) differs from the past published ones in that the macromolecular envelope is a homographic exponential model and therefore can be parametrized with a few coefficients of ω while still retaining its 'binary-like' character.

This work was partially supported by a TMR Marie Curie Grant (to PR) and a Sponsored Research Agreement from Pfizer Central Research (to GB). We wish to thank one of the referees for extremely helpful reviewing of the manuscript.

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.
Abrahams, J. P. & Leslie, A. (1996). *Acta Cryst.* D**52**, 30–42.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.
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.
Guo, D., Blessing, R. H. & Langs, D. A. (2000). *Acta Cryst.* D**56**, 451–457.
Leslie, A. (1987). *Acta Cryst.* A**43**, 134–136.
Meyer, E. F., Radhakrishnan, R., Cole, G. M. & Presta, L. G. (1986). *J. Mol. Biol.* **189**, 553–559.
Perrakis, A., Morris, R. & Lamzin, V. (1999). *Nature Struct. Biol.* **6**(2), 458–463.
Ramin, M. (1999). PhD thesis. LURE, Université Paris XI, Orsay, France.
Reif, F. (1965). *Fundamentals of Statistical and Thermal Physics*, 1st ed., pp. 350–351. Singapore: McGraw–Hill.
Roversi, P., Irwin, J. & Bricogne, G. (1998). *Acta Cryst.* A**54**, 971–996.
Roversi, P., Irwin, J. & Bricogne, G. (2000). In *Electron, Spin and Momentum Densities and Chemical Reactivities*, edited by P. G. Mezey & B. E. Robertson. Dordrecht: Kluwer. In the press.
Schiltz, M., Shepard, W., Fourme, R., Prangé, T., de La Fortelle, E. & Bricogne, G. (1997). *Acta Cryst.* D**53**, 78–92.
Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1863–1871.
Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.
Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.
Wang, B.-C. (1985). *Methods Enzymol.* **112**, 813–815.