Acta Crystallographica Section D
**Biological Crystallography**
ISSN 1399-0047

# Metrics for comparison of crystallographic maps

**Alexandre Urzhumtsev,[a,b]* Pavel V. Afonine,[c] Vladimir Y. Lunin,[d] Thomas C. Terwilliger[e] and Paul D. Adams[c,f]**

[a]Centre for Integrative Biology, Department of Integrated Structural Biology, IGMBC, CNRS UMR 7104–INSERM U964–Université de Strasbourg, 1 Rue Laurent Fries, BP 10142, 67404 Illkirch, France, [b]Faculté des Sciences et Technologies, Université de Lorraine, 54506 Vandoeuvre-lès-Nancy, France, [c]Lawrence Berkeley National Laboratory, One Cyclotron Road, BLDG 64R0121, Berkeley, CA 94720, USA, [d]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino 142290, Russian Federation, [e]Los Alamos National Laboratory, Los Alamos, NM 87545-0001, USA, and [f]Department of Bioengineering, University of California Berkeley, Berkeley, CA 94720, USA

Correspondence e-mail: sacha@igbmc.fr

Received 21 February 2014
Accepted 14 July 2014

Numerical comparison of crystallographic contour maps is used extensively in structure solution and model refinement, analysis and validation. However, traditional metrics such as the map correlation coefficient (map CC, real-space CC or RSCC) sometimes contradict the results of visual assessment of the corresponding maps. This article explains such apparent contradictions and suggests new metrics and tools to compare crystallographic contour maps. The key to the new methods is rank scaling of the Fourier syntheses. The new metrics are complementary to the usual map CC and can be more helpful in map comparison, in particular when only some of their aspects, such as regions of high density, are of interest.

## 1. Notation

$F(hkl)\exp[i\varphi(hkl)]$: crystallographic structure factor with indices $hkl$.
$\mathbf{F}_{\rm calc} = F_{\rm calc}\exp(i\varphi_{\rm calc})$: structure factors calculated from an atomic model.
$\mathbf{F}_{\rm model} = F_{\rm model}\exp(i\varphi_{\rm model})$: structure factors calculated from an atomic model including modelled contribution from bulk solvent and various scales (Afonine *et al.*, 2013).
$N_x$, $N_y$, $N_z$: grid numbers defining a regular grid in real space.
$N_{\rm grid}$: total number of grid nodes of the unit cell used for comparison; in particular, $N_{\rm grid} = N_x \times N_y \times N_z$ if the maps are analyzed for the whole unit cell.
$\mathbf{n} = (n_x, n_y, n_z)$: grid node defined by its three integer indices.
$\rho(x, y, z)$: Fourier synthesis calculated in the unit cell of direct space.
$\rho(\mathbf{n}) = \rho(n_x, n_y, n_z)$: Fourier synthesis calculated in grid node $\mathbf{n}$.
$\rho_\sigma(\mathbf{n}) = \rho_\sigma(n_x, n_y, n_z)$: Fourier synthesis scaled in $\sigma$.
$\rho_{d1-d2}$: Fourier synthesis calculated with structure factors in the resolution range $(d_1, d_2)$.
$\rho_{\rm complete}$, $\rho_{\rm incomplete}$: Fourier syntheses calculated with a complete set of structure factors up to a given high-resolution cutoff or with some reflections excluded from this set; both the resolution value and the method used to exclude reflections are described explicitly for particular tests.
$(F, \varphi)$ synthesis: Fourier synthesis calculated with the Fourier coefficients $F\exp(i\varphi)$.
$N_\mu$: number of grid nodes with the value below the cutoff level $\mu$ in the Fourier synthesis $\rho$: $\rho(\mathbf{n}) < \mu$; $\mu$ is given in the same units as $\rho$.
$\eta(\mu; \rho)$: quantile rank corresponding to the cutoff level $\mu$ for the Fourier synthesis $\rho(\mathbf{n})$.
$Q(\mathbf{n})$: (quantile) rank-scaled Fourier synthesis $\rho(\mathbf{n})$.
$P(\mathbf{n})$: rank-scaled Fourier synthesis $\rho(\mathbf{n})$ with the values flattened out of the peaks.
$M(q) = \{\mathbf{n}: Q(\mathbf{n}) < q\}$: mask defined by the cutoff level expressed in the quantile rank $q$.

$D(q; \rho_a, \rho_b)$; discrepancy function between two grid functions, $\rho_a(\mathbf{n})$ and $\rho_b(\mathbf{n})$, in particular between two Fourier syntheses.
$CC(\rho_a, \rho_b)$: map correlation coefficient between two grid functions.
$CC_r(\rho_a, \rho_b)$: rank correlation coefficient between two grid functions.
$CC_{<q\text{peak}>}(\rho_a, \rho_b)$: peak correlation coefficient between two grid functions; selected peaks correspond to the $q_{\text{peak}}$ quantile rank.

## 2. Introduction

Macromolecular crystallography operates with the electron (or neutron) density distribution in crystals. For ideal crystals, this physical entity can be described by a periodic function $\rho_{\text{exact}}(x, y, z)$ of three space fractional coordinates $(x, y, z)$ and can be represented by a Fourier series composed of an infinite number of complex coefficients $F(hkl)\exp[i\varphi(hkl)]$,

$$\rho_{\text{exact}}(x, y, z) = \kappa \sum_{hkl=-\infty}^{\infty} F(hkl)\exp[i\varphi(hkl)]\exp[-2\pi i(hx + ky + lz)] \quad (1)$$

(Ewald, 1913). The values of these coefficients, called structure factors, depend on the crystal under study. The scale factor $\kappa$, equal to the inverse unit-cell volume, puts function (1) on an absolute scale; alternative scales can also be used. In crystallographic practice, Fourier series contain only a finite set $S$ of terms and are usually calculated on a three-dimensional regular grid $N_x \times N_y \times N_z$ with the grid nodes described by integer indices $\mathbf{n} = (n_x, n_y, n_z)$,

$$\rho(\mathbf{n}) = \rho(n_x, n_y, n_z) = \kappa \sum_{hkl\in S} F(hkl)\exp[i\varphi(hkl)] \times \exp\left[-2\pi i\left(h\frac{n_x}{N_x} + k\frac{n_y}{N_y} + l\frac{n_z}{N_z}\right)\right]. \quad (2)$$

We call these grid functions (2) Fourier syntheses. To be analyzed visually or by a computer program, these mathematical entities are traditionally explored by contouring three-dimensional isosurfaces

$$\rho(\mathbf{n}) = \mu_1, \rho(\mathbf{n}) = \mu_2, \ldots, \quad (3)$$

where $\mu_n$ are empirically chosen values. The result of such contouring is a geometric object that is referred to below as a crystallographic contour map.

Crystallographic structure solution typically deals with many maps arising at different stages of the process. Often, one is required to compare maps in order to assess model-building and/or refinement steps. Quantitative comparison of maps calculated for the same crystal, for different crystals and even for different structures is important to evaluate the progress of structure solution and to validate the structure. However, confusion about the three terms given above, electron (or neutron) density distribution, Fourier syntheses and corresponding Fourier contour maps, sometimes leads to apparent contradictions between numerical and visual analyses, as shown below.

As an example, we consider the exact electron density $\rho_{\text{pept\_}a}(\mathbf{n}) = \rho_{\text{exact}}(\mathbf{n})$ corresponding to a peptide model ($B$ = 1 Å$^2$) placed in an orthogonal unit cell with unit-cell parameters $a = b = 6$, $c = 3$ Å, space group $P1$. $\rho_{\text{pept\_}b}(\mathbf{n})$ is its Fourier synthesis at a resolution of 0.5 Å and $\rho_{\text{pept\_}c}(\mathbf{n})$ is a Fourier synthesis calculated at a resolution of 1.0 Å for the same peptide model but taken with $B = 5$ Å$^2$ and completed by a water molecule with $B = 20$ Å$^2$.

The maps for $\rho_{\text{pept\_}a}(\mathbf{n})$ and $\rho_{\text{pept\_}b}(\mathbf{n})$ shown at $2\sigma$ (§3.1.1) are very similar to each other (compare Fig. 1*a* with Fig. 1*b*). However, the usual map correlation coefficient

$$CC(\rho_a, \rho_b) = \frac{\sum_{\mathbf{n}}[\rho_a(\mathbf{n}) - \langle\rho_a\rangle][\rho_b(\mathbf{n}) - \langle\rho_b\rangle]}{\left\{\sum_{\mathbf{n}}[\rho_a(\mathbf{n}) - \langle\rho_a\rangle]^2\right\}^{1/2}\left\{\sum_{\mathbf{n}}[\rho_b(\mathbf{n}) - \langle\rho_b\rangle]^2\right\}^{1/2}} \quad (4)$$

(see Supporting Information[1] §S1) between $\rho_a(\mathbf{n}) = \rho_{\text{pept\_}a}(\mathbf{n})$ and $\rho_b(\mathbf{n}) = \rho_{\text{pept\_}b}(\mathbf{n})$ is only 0.90; here, $\langle\rho_a\rangle$ and $\langle\rho_b\rangle$ represent the mean values of $\rho_a(\mathbf{n})$ and $\rho_b(\mathbf{n})$, respectively. Indeed, the contour maps at $1\sigma$ (compare Fig. 1*d* with Fig. 1*e*) show that $\rho_{\text{pept\_}b}(\mathbf{n})$ differs significantly from $\rho_{\text{pept\_}a}(\mathbf{n})$. This reminds us that similarity of two contour maps at some cutoff level does not necessarily imply similarity of the corresponding syntheses.

Note that here we use the coefficient (4) to compare the whole syntheses, for example as in Read (1986) and Lunin & Woolfson (1993), while it can also be used locally (see, for example, Brändén & Jones, 1990; Kleywegt *et al.*, 2004; Rupp, 2006; Tickle, 2012).

Secondly, the traditional choice of a cutoff level in $\sigma$ (§3.1.1) is often not appropriate for map comparison. The map for $\rho_{\text{pept\_}c}(\mathbf{n})$ at $2\sigma$ (Fig. 1*f*) shows a much larger volume of the unit cell in comparison with that for $\rho_{\text{pept\_}a}(\mathbf{n})$ at the same $2\sigma$ cutoff level (Fig. 1*a*). However, the maps look similar when taken at different cutoff values (compare Fig. 1*c* with Fig. 1*a*).

Thirdly, the three maps $\rho_{\text{pept\_}a}(\mathbf{n})$, $\rho_{\text{pept\_}b}(\mathbf{n})$ and $\rho_{\text{pept\_}c}(\mathbf{n})$ look similar to each other, while the map correlation coefficient CC calculated using (4) is high for one pair of them, $CC(\rho_{\text{pept\_}a}, \rho_{\text{pept\_}b}) = 0.9$, and is low for another, $CC(\rho_{\text{pept\_}a}, \rho_{\text{pept\_}c}) = 0.6$.

In fact, the map correlation coefficient (4) is obtained by comparing two sets of values calculated on the same grid, comparing all these values point by point but with no reference to the position of these points in space (these may even be in a one-dimensional space). However, when we compare two maps visually we look at the shape of one or a few chosen isosurfaces. In other words, these two methods of comparison give different characteristics for different objects related to each other as explained above.

Fig. 2 illustrates a practical example with two protein models available in the PDB (Bernstein *et al.*, 1977; Berman *et*

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: KW5094).

*al.*, 2000). Here again, the calculated CC values disagree with the visual analysis. The corresponding details are given in §4.2.1.

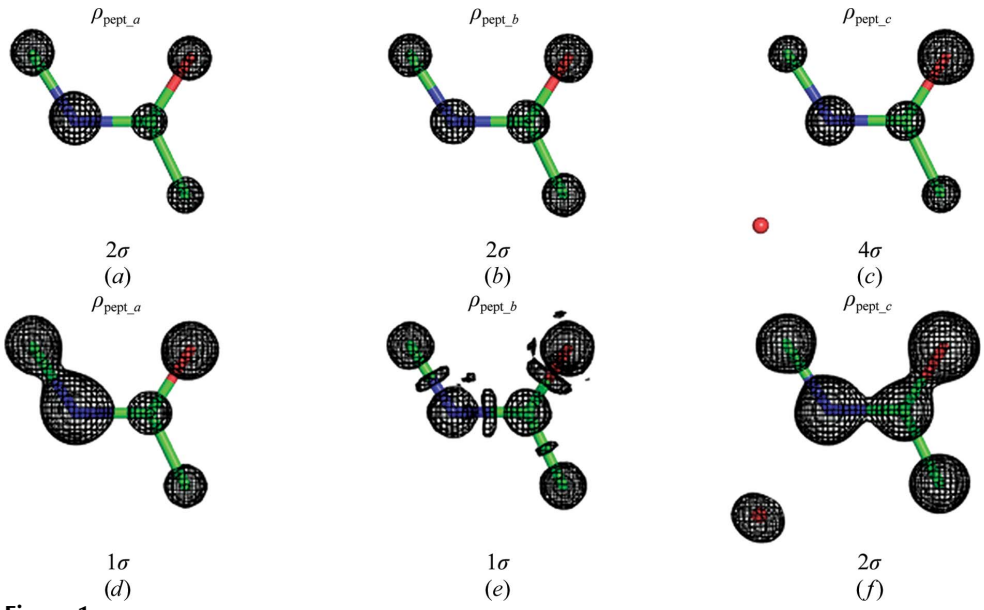


**Figure 1**
Fourier contour maps for artificial crystallographic peptide data. The function $\rho_{pept\_a}(\mathbf{n})$ is an exact electron-density distribution for the peptide model with $B = 1$ Å$^2$; $\rho_{pept\_b}(\mathbf{n})$ is the corresponding Fourier synthesis at a resolution of 0.5 Å. $\rho_{pept\_c}(\mathbf{n})$ is the Fourier synthesis at a resolution of 1.0 Å for the same model completed by a water molecule and taken with $B = 5$ Å$^2$. All H atoms were excluded from the calculations.

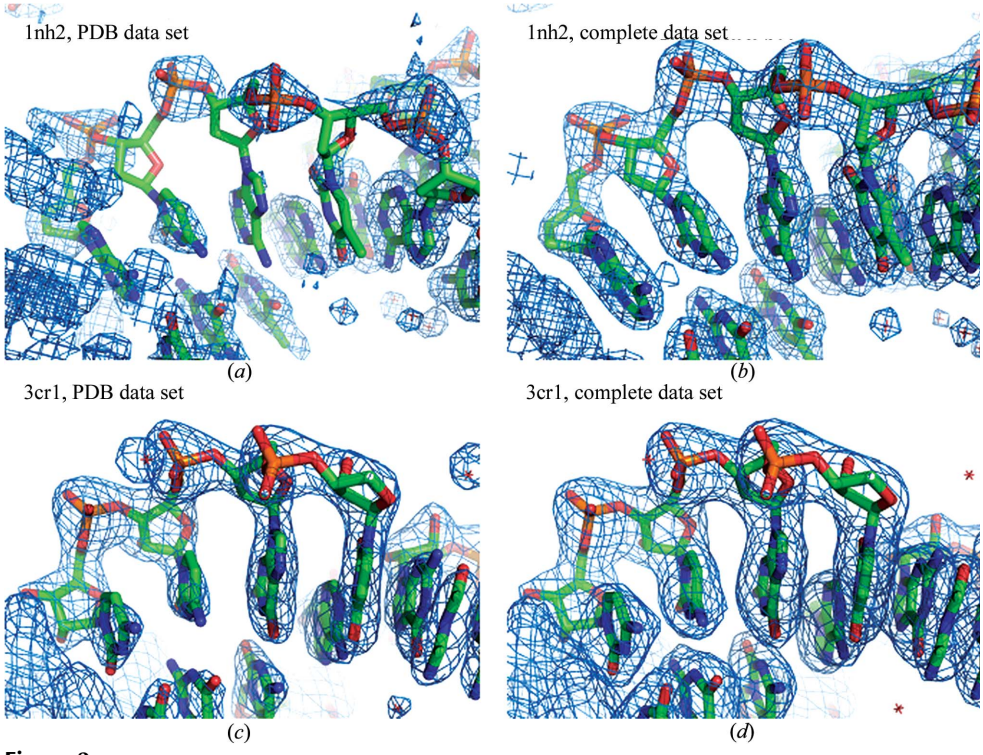


**Figure 2**
Fourier contour maps for 1nh2 and 3cr1. All syntheses are calculated with the model structure factors at a resolution of 1.90 Å (1nh2) or 2.25 Å (3cr1). The syntheses are obtained for the complete data sets (right column) and for those from the PDB (left column). The 1nh2 maps are shown with a cutoff level of 1.0$\sigma$ and those for 3cr1 with a cutoff level of 1.5$\sigma$. See §4.2 for details. The map correlation coefficient between the top syntheses is 0.702 and that between the bottom syntheses is 0.642.

Crystallographers use contour maps at different contour levels to focus on different aspects of the maps. At high contour levels the most prominent features are shown, while at lower contour levels more details of the electron density are seen. In many cases it is the most prominent features that are most useful to the crystallographer in identifying where atoms are likely to be present in the structure. In other cases, of course, the details of the map are very important in identifying errors in atomic placement and in comparing different maps.

In this article, we focus on a subset of the information in a map, such as the prominent features in the electron density, and suggest new approaches to comparing crystallographic maps. The emphasis in this work is on the shapes of isosurfaces in these maps. These are the shapes that crystallographers normally use to identify the atomic features of structures in crystals.

Suppose we have two functions calculated on the same grid. For each function a mask can be defined by some isosurface, with all the points inside this mask having a value greater than the cutoff associated with the isosurface. We would like to compare the shapes of these masks (isosurfaces). Intuitively, masks containing a different number of grid nodes are different. The question we focus on is how similar are two masks composed of the same number of grid nodes, *i.e.* covering the same volume of the unit cell. We show below that to answer this question it is convenient to rescale the syntheses in the quantile rank (see §3.1.2) instead of a traditional scaling in $\sigma$ (see §3.1.1).

After introducing rank scaling, we discuss a way to create a normalized metric useful in the comparison of two masks or a series of masks for various cutoff levels (§3.2). This naturally leads to a use of the Spearman rank correlation (Spearman, 1904; see

also, for example, Lehmann & D'Abrera, 1998 and references therein), which is the same as the conventional correlation coefficient calculated for rank-scaled maps (§3.3). Considering only grid nodes with relatively high rank values results in another metric, a peak correlation coefficient (§3.4) that corresponds to a visual comparison of the contour maps and that is based on much of the key structural information in the maps. §4 gives various possible illustrations where the new metrics complement the traditional map correlation coefficient or explain some its apparent contradiction with a visual analysis.

Comparison of maps calculated on different grids is outside the scope of this work.

## 3. Methods

### 3.1. Scaling of crystallographic Fourier syntheses

**3.1.1. Scaling by $\sigma$.** In macromolecular crystallography, currently the most popular way of scaling crystallographic syntheses is by $\sigma$. Sigma-scaled Fourier syntheses are obtained as follows,

$$\rho_\sigma(\mathbf{n}) = \frac{1}{\sigma_\rho}[\rho(\mathbf{n}) - \langle\rho\rangle] \tag{5}$$

with

$$\langle\rho\rangle = \frac{\sum\limits_{\mathbf{n}} \rho(\mathbf{n})}{\sum\limits_{\mathbf{n}} 1} = \frac{\sum\limits_{\mathbf{n}} \rho(\mathbf{n})}{N_{\rm grid}} \tag{6}$$

and

$$\sigma_\rho = \left\{ \frac{\sum\limits_{\mathbf{n}} [\rho(\mathbf{n}) - \langle\rho\rangle]^2}{N_{\rm grid}} \right\}^{1/2}. \tag{7}$$

Here, $\rho(\mathbf{n})$ is some initial function, $N_{\rm grid}$ is the number of grid points in the unit cell and $\langle\rho\rangle$ is always equal to 0 when the term $F_{000}$ is absent from the Fourier series (2). With such a scaling, the grid function (5) has the properties

$$\sum_{\mathbf{n}} \rho_\sigma(\mathbf{n}) = 0 \tag{8}$$

and

$$\left[ \frac{\sum\limits_{\mathbf{n}} \rho_\sigma^2(\mathbf{n})}{N_{\rm grid}} \right]^{1/2} = 1. \tag{9}$$

Empirically, crystallographers consider values of $\rho_\sigma(\mathbf{n}) > 1$ as a 'signal level' at which the structural details are analyzed (values notably above the mean value, *i.e.* above the value for bulk solvent) and values of $\rho_\sigma(\mathbf{n}) > 3$ as a 'strong signal level'.

Another source of confusion comes from the map correlation coefficient (4). In statistics, the correlation coefficient is used to compare two sets of values from related distributions. However, the same formal expression is often used in crystallography, instead of the least-squares metric (Supporting Information §S1), to compare two syntheses defined as vectors in an $N_{\rm grid}$-dimensional space. We stress that in the current work we do not consider the crystallographic Fourier syntheses as random functions even when such a consideration has previously been used in a number of projects (see, for example, Luzzati, 1953; Blow & Crick, 1959; Ramachandran & Raman, 1959; Main, 1979 and references therein; Vijayan, 1980; Read, 1986; Lunin, 1989; Terwilliger, 2000; Burla *et al.*, 2010; Lang *et al.*, 2014). In the following, we consider that both the map correlation coefficient (4) and the new metrics are calculated for the whole unit cell. Naturally, they can be calculated locally for any part of the unit cell; in this case, $N_{\rm grid}$ would be the number of grid nodes inside this part.

Since the scaling (5)–(7) is a linear transformation, the correlation coefficient (4) calculated for the $\rho_\sigma(\mathbf{n})$ values coincides with the correlation coefficient CC calculated using the original values $\rho(\mathbf{n})$.

While such scaling in $\sigma$ is convenient to distinguish macromolecular features, it may be misleading when used for visual and numerical comparison of syntheses, as the example in §2 shows (Fig. 1; see also §4.1). The reason for this is that the frequency distribution of the values of the syntheses (Lunin, 1988, 1993; Main, 1990*a*,*b*) may be different for the two syntheses. As a consequence, the same cutoff level in $\sigma$ defines different numbers of grid nodes selected by this level for these syntheses. Obviously, regions composed of a different number of points (using the same grid) can never be equal.

**3.1.2. Rank scaling**. The map comparison becomes easier if the Fourier syntheses are scaled in quantile ranks or are rank scaled. In image processing, this operation is referred to as histogram equalization (see, for example, Pratt, 1978). This means that for each cutoff value $\mu$ we count the number $N_\mu$ of grid nodes $\mathbf{n}$ such that the synthesis value is below it, $\rho(\mathbf{n}) < \mu$, and we then calculate the ratio

$$\eta(\mu;\,\rho) = \frac{N_\mu}{N_{\rm grid}}, \quad 0 \le \eta(\mu;\,\rho) \le 1. \tag{10}$$

Here, the second argument, $\rho$, is the Fourier synthesis to be studied and the first argument, $\mu$, is a particular value. In statistics, the value $\eta$ (10) is called a quantile rank; when multiplied by 100 this gives the percentile rank. The notions of percentile and quantile and the corresponding ranks have recently been used in crystallography by Pozharski (2010), Gore *et al.* (2012) and Tickle (2012), although for different goals. Previously in crystallography, a scaling in units complementary to the quantile/percentile rank, *i.e.* in the fractional unit-cell volume covered by the mask $\rho(\mathbf{n}) > \mu$, has been used by Vagin (personal communication) and by Lunin and coworkers (Lunin, 1988; Vernoslova & Lunin, 1993).

For a given synthesis $\rho$, the function (10) increases with $\mu$. This monotonic behaviour permits an easy rank scaling (Appendix *A*), replacing the value $\rho(\mathbf{n})$ at each point by

$$Q(\mathbf{n}) = \eta[\rho(\mathbf{n});\,\rho] \tag{11}$$

using $\eta(\mu;\,\rho)$ (10). This scaling does not change the shape of any isosurface, as all points with the same value of $\mu$ have the

same value of the new function. Note that in contrast to the rescaling in $\sigma$, rank rescaling is a nonlinear transformation.

Most commonly, macromolecular crystallographers work with syntheses calculated with the coefficients (amplitudes) $wF_{\rm obs}$ or $2mF_{\rm obs} - DF_{\rm calc}$ (Read, 1986) and scaled in $\sigma$. Analyzing these syntheses, at least for the resolutions 1–3 Å at which many structural projects are carried out, the cutoff values $\mu$ used for visual interpretation range approximately between 1 and $2\sigma$. The particular choice may depend on the resolution, bulk-solvent content and other factors. Fig. 3 shows that the ranks corresponding to these values vary approximately from 0.85 to 0.95. These model calculations agree with calculations using various experimental data (not shown); in particular, this includes experimental data from PDB entries with low, medium and extremely high solvent content. This

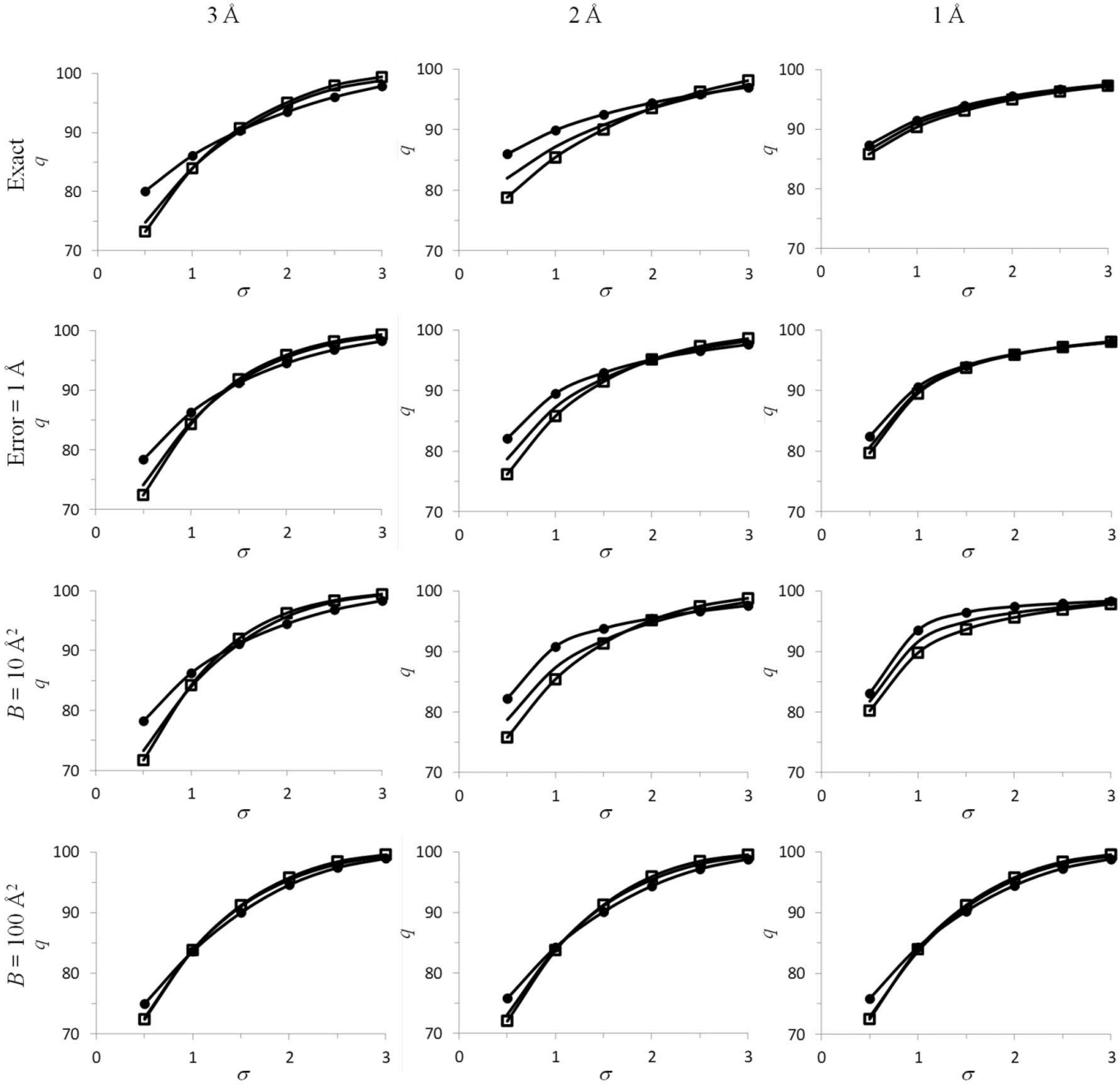


**Figure 3**
Quantile ranks (multiplied by 100) corresponding to a different $\sigma$ cutoff in Fourier syntheses at resolutions of 3 Å (left column), 2 Å (central column) and 1 Å (right column). Syntheses are computed with the exact structure factors calculated from an accurate atomic model (top row) and from a model with large random coordinate errors (second row) with a small (third row) and a large (bottom row) atomic displacement parameter ($B$). The model was placed in unit cells of different sizes simulating different percentages of bulk-solvent content equal to 0.23 (empty square markers), 0.58 (no markers) and 0.81 (solid circle markers).

also agrees with the previous observation by Ioerger & Sacchettini (2002).

Other scaling methods, *e.g.* choosing another $\kappa$ value [for example such that $\max_{\mathbf{n}}|\rho(\mathbf{n})| = 100$ or using a so-called 'absolute scale'] or another nonlinear scheme (for example, Bhat, 1988; Lunin *et al.*, 2000) are known, but we will not review this issue here in detail.

### 3.2. Comparison of two masks

Since the introduction of graphics stations in macromolecular crystallography, syntheses have typically been presented by a single isosurface at a time with the possibility of varying the corresponding cutoff levels. When we compare two syntheses visually, we look at the shape of the masks covered by the corresponding isosurfaces (there are a number of publications on image analysis that discuss the relevant computational procedures; see, for example, Bruckner & Möller, 2010 and references therein). As mentioned above, the quantitative similarity of two masks can be examined most readily when these masks are constructed so that they contain equal volumes. This is a particular advantage of the rank-scaling approach, which naturally leads to equal volumes at given contour levels in different maps. Other one-to-one syntheses-scaling schemes with a similar property (Supporting Information §S3) are less convenient for the goals of the current work.

In order to compare two masks, we start by measuring (calculating) the difference between them. Let $Q_a(\mathbf{n})$ and $Q_b(\mathbf{n})$ be the two rank-rescaled syntheses $\rho_a(\mathbf{n})$ and $\rho_b(\mathbf{n})$. For any quantile rank value, $0 \leq q \leq 1$, the subsets (masks)

$$M_a(q) = \{\mathbf{n} : Q_a(\mathbf{n}) < q\}$$
$$M_b(q) = \{\mathbf{n} : Q_b(\mathbf{n}) < q\} \quad (12)$$

contain the same number $N_{\rm selected} = qN_{\rm grid}$ of grid nodes. The difference between these masks may be described by the number $N_{\rm diff}$ of the nodes that belong to one of them and do not belong to another one,

$$M_{\rm diff}(q) = M_{ab}(q) \cup M_{ba}(q), \quad (13)$$

where

$$M_{ab}(q) = \{\mathbf{n} : [Q_a(\mathbf{n}) < q] \text{ and } [Q_b(\mathbf{n}) \geq q]\}$$
$$M_{ba}(q) = \{\mathbf{n} : [Q_a(\mathbf{n}) \geq q] \text{ and } [Q_b(\mathbf{n}) < q]\}. \quad (14)$$

Note that by construction $M_{ab}(q)$ and $M_{ba}(q)$ contain the same number of points. The condition $N_{\rm diff} = 0$ means that the masks $M_a(q)$ and $M_b(q)$ coincide. If $N_{\rm diff} > 0$ then the masks are different, but the value of $N_{\rm diff}$ does not allow judgment of the degree of this difference because the same number of differing points $N_{\rm diff}$ may have a different significance for small and large rank values $q$.

To put this difference $N_{\rm diff}$ on a scale, we compare $M_a(q)$ with a random set $M_{\rm random}$ composed of the same number $N_{\rm selected}$ of grid nodes distributed uniformly in the unit cell and thus containing no structural information. On average, the number of grid nodes of $M_{\rm random}$ that are outside $M_a(q)$ is just the number of grid nodes in $M_{\rm random}$ multiplied by the fraction of the cell that is outside $M_a(q)$, *i.e.* by $(1 - q)$,

$$(1 - q)N_{\rm selected} = (1 - q)qN_{\rm grid}. \quad (15)$$

The same estimate is valid for the comparison of $M_{\rm random}$ with $M_b(q)$. Based on this, we normalize $N_{\rm diff}$ as

$$D(q; \rho_a, \rho_b) = \frac{N_{\rm diff}}{2q(1 - q)N_{\rm grid}}. \quad (16)$$

The calculated values of the normalized function $D(q; \rho_a, \rho_b)$ at some value of the argument $q$ may be equal to its minimal possible value, zero, when the corresponding masks coincide and may approach one when the two masks are uncorrelated. Since (15) is only a statistical estimate, in practice $D(q; \rho_a, \rho_b)$ may sometimes happen to be greater than one. We notate (16) as $D(q; \rho_a, \rho_b)$ and not $D(q; Q_a, Q_b)$ to stress that this measure can be applied to any two functions calculated on the same grid and not necessarily functions rescaled in some specific way. We call (16) a discrepancy function. Different values of the argument $q$ are useful for obtaining different types of information: high $q$ values are useful for identifying the peaks of the functions (atomic positions or macromolecular chain), while $q$ close to 0.5 is useful for the identification of molecular envelopes (the actual corresponding value of $q$ varies with the fraction of the solvent region).

### 3.3. Rank correlation coefficient

When calculating the discrepancy function $D$ (16) between two syntheses, we compare masks of equal size ('equivalent masks'), varying the cutoff level at which these masks are selected. To make such comparison easier, we rank-scale the syntheses. When comparing a pair of equivalent masks we check each grid node one by one, identifying whether this grid node is inside only one mask, the other, both or neither.

Alternatively, after rank scaling the two syntheses $Q_a(\mathbf{n})$ and $Q_b(\mathbf{n})$ we may express their similarity by

$$\mathrm{CC_r}(\rho_a, \rho_b) = \frac{\sum_{\mathbf{n}}[Q_a(\mathbf{n}) - \langle Q_a\rangle][Q_b(\mathbf{n}) - \langle Q_b\rangle]}{\left\{\sum_{\mathbf{n}}[Q_a(\mathbf{n}) - \langle Q_a\rangle]^2\right\}^{1/2}\left\{\sum_{\mathbf{n}}[Q_b(\mathbf{n}) - \langle Q_b\rangle]^2\right\}^{1/2}}. \quad (17)$$

This metric of similarity of syntheses $\rho_a(\mathbf{n})$ and $\rho_b(\mathbf{n})$ varies from $-1$ to 1, and in statistics it is known as Spearman's rank correlation coefficient (Spearman, 1904). We may note that (Appendix *A*)

$$\mathrm{CC_r}(\rho_a, \rho_b) \simeq \frac{12\sum_{\mathbf{n}} Q_a(\mathbf{n})Q_b(\mathbf{n})}{N_{\rm grid}} - 3. \quad (18)$$

The key property of the rank correlation coefficient $\mathrm{CC_r}(\rho_a, \rho_b)$ is its invariance with respect to scaling of the syntheses $\rho_a(\mathbf{n})$ and $\rho_b(\mathbf{n})$. As mentioned in §3.1.1, scaling by $\sigma$ does not change the standard correlation coefficient $\mathrm{CC}(\rho_a, \rho_b)$. In particular, $\mathrm{CC}(\rho_a, \rho_b) = 1$ for all proportional functions, *i.e.* when $\rho_b(\mathbf{n}) = \lambda\rho_a(\mathbf{n})$ for all $\mathbf{n}$. An important advantage of the rank correlation coefficient $\mathrm{CC_r}(\rho_a, \rho_b)$ compared with $\mathrm{CC}(\rho_a,$

$\rho_b$) is that the former is invariant upon any monotonic (and not necessary linear) rescaling of the syntheses $\rho_a(\mathbf{n})$ and $\rho_b(\mathbf{n})$. In particular, $CC_r(\rho_a, \rho_b)$ for a pair of nonproportional functions related by any monotonously increasing function $\rho_b(\mathbf{n}) = f[\rho_a(\mathbf{n})]$.

Note that using $CC_r(\rho_a, \rho_b)$, in contrast to $D(q; \rho_a, \rho_b)$, applies not only to Fourier maps shown as series of masks but also to any continuous spectrum of colours or intensities (see, for example, Schotte *et al.*, 2003).

As an example, the rank correlation coefficients $CC_r$ for the peptide syntheses defined in §2 are given by $CC_r(\rho_{pept\_a}, \rho_{pept\_b}) = 0.56$ and $CC_r(\rho_{pept\_a}, \rho_{pept\_b}) = 0.22$, which is more indicative of their difference than the standard map correlation coefficient values, which are equal to 0.90 and 0.60, respectively. More details of comparison of these syntheses using the discrepancy function, the rank correlation coefficient and other metrics as defined below are discussed in §4.1.

### 3.4. Comparison of peaks

Syntheses such as $wF_{obs}$ or $2mF_{obs} - DF_{calc}$ scaled in $\sigma$ have both positive and negative values. While analysis of negative values may be important (see, for example, Urzhumtsev *et al.*, 1989), often only the regions of positive values are of interest. This is the case for visual analysis and manual model building; for example, the program *Coot* (Emsley *et al.*, 2010) defaults to showing nondifference $\sigma$-scaled maps at $\mu > 0$. However, maps similar in the positive domain may be different in the negative domain. This may give rise to an apparent contradiction: similar-looking maps (inspected in the positive domain only) may have low correlations computed using the entirety of the maps.

Since map regions with high values contain most of the structural information, it is useful to have a way to compare contour maps such that (i) differences between low values of the synthesis should not play a role and (ii) if a high value in one synthesis corresponds to a low value in another synthesis, the desired metric should not depend on the exact value of the lower value.

For example, for structures with the most frequent percentage of bulk solvent, the separation of positive and negative values in $\sigma$-scaled maps roughly corresponds to half of the syntheses, *i.e.* to the quantile-rank cutoff $q = 0.50$. When comparing the top halves of the rank-scaled syntheses $Q_a(\mathbf{n})$ and $Q_b(\mathbf{n})$, we shall exclude from comparison all grid points for which the values in both syntheses are low, defining a set of grid nodes staying with

$$\Omega_{50} = \{\mathbf{n} : [Q_a(\mathbf{n}) > 0.50] \text{ or } [Q_b(\mathbf{n}) > 0.50]\}. \quad (19)$$

Similarly, to effectively compare regions with high density (near peaks in the map) corresponding to a higher quantile rank value $0.5 < q_{peak} < 1.0$, we define

$$\Omega_{q\text{peak}} = \{\mathbf{n} : [Q_a(\mathbf{n}) > q_{peak}] \text{ or } [Q_b(\mathbf{n}) > q_{peak}]\}. \quad (20)$$

We then flatten the syntheses values in the $\Omega_{q\text{peak}}$ points if these values are below $q_{peak}$ for one of the syntheses,

$$P_a(\mathbf{n}) = \begin{cases} Q_a(\mathbf{n}) & \text{if } Q_a(\mathbf{n}) \geq q_{peak} \\ q_{peak} & \text{if } Q_a(\mathbf{n}) < q_{peak} \end{cases},$$
$$P_b(\mathbf{n}) = \begin{cases} Q_b(\mathbf{n}) & \text{if } Q_b(\mathbf{n}) \geq q_{peak} \\ q_{peak} & \text{if } Q_b(\mathbf{n}) < q_{peak} \end{cases} \quad (21)$$

and finally calculate

$$CC_{<q\text{peak}>}(\rho_a, \rho_b) = \frac{\sum_{\mathbf{n}\in\Omega_{q\text{peak}}} [P_a(\mathbf{n}) - \langle P_a\rangle][P_b(\mathbf{n}) - \langle P_b\rangle]}{\left\{\sum_{\mathbf{n}\in\Omega_{q\text{peak}}} [P_a(\mathbf{n}) - \langle P_a\rangle]^2\right\}^{1/2}\left\{\sum_{\mathbf{n}\in\Omega_{q\text{peak}}} [P_b(\mathbf{n}) - \langle P_b\rangle]^2\right\}^{1/2}} \quad (22)$$

Here,

$$\langle P_a\rangle = \frac{\sum_{\mathbf{n}\in\Omega_{q\text{peak}}} P_a(\mathbf{n})}{N_{\Omega,q\text{peak}}}, \quad \langle P_b\rangle = \frac{\sum_{\mathbf{n}\in\Omega_{q\text{peak}}} P_b(\mathbf{n})}{N_{\Omega,q\text{peak}}} \quad (23)$$

and $N_{\Omega,q\text{peak}}$ is the number of grid nodes in $\Omega_{q\text{peak}}$ defined by (20). For example, a $q_{peak}$ value equal to 0.50 defines $CC_{50}$ and a $q_{peak}$ value equal to 0.90 defines $CC_{90}$ . As previously, the sums in (22) exclude all grid nodes in which both syntheses have values lower than the chosen threshold, indicating that we are not interested in comparison of syntheses at these points.

### 3.5. Practical applications

Depending on the particular problem, different tools are useful to compare crystallographic Fourier syntheses and the corresponding contour maps.

Naturally, when the similarity of three-dimensional functions (for example, crystallographic Fourier syntheses) is analyzed, for example when these functions are used to extract the phase values of corresponding Fourier coefficients, the traditional map correlation coefficient (4) is still a good metric.

However, in a major part of crystallographic projects only the Fourier contour maps for positive cutoff values (in $\sigma$-scaled syntheses) are used for visual inspection of maps. Moreover, for syntheses at a resolution of 1–3 Å the most frequently used cutoff levels of 1–2$\sigma$ correspond to rank values $q$ of as high as 0.85–0.95. To accompany the traditional visual analysis, we suggest using the coefficient $CC_{90}$ as a rule of thumb and switching to $CC_{95}$ using higher rank values in the case of a larger fraction of bulk solvent, higher map resolution or smaller $B$ factors, and switching to $CC_{85}$ or $CC_{80}$ in the opposite situations. The correlation coefficient $CC_{50}$ may be used to characterize the similarity of isosurfaces roughly corresponding to molecular masks for structures with typical values of the bulk-solvent fraction.

Use of the coefficient $CC_r$ may be advised when the whole set of isosurfaces, including those for negative peaks, are studied. The discrepancy function $D(q; \rho_1, \rho_2)$ completes this toolset when more detailed information is required.

§4 below provides examples of applications of the new correlation coefficients to macromolecular diffraction data. All of these applications confirm that the new metrics reflect

**Table 1**
Numerical comparison of the syntheses for the peptide model.

For definition of the syntheses (Fig. 1) and the correlation coefficients between $\rho_a$ and $\rho_b$, see the text.

| $\rho_a$ | $\rho_b$ | CC | $CC_r$ | $CC_{50}$ | $CC_{70}$ | $CC_{80}$ | $CC_{90}$ | $CC_{95}$ | $CC_{99}$ |
|---|---|---|---|---|---|---|---|---|---|
| $\rho_{pept\_a}$ | $\rho_{pept\_b}$ | 0.895 | 0.557 | 0.597 | 0.663 | 0.708 | 0.485 | 0.517 | 0.829 |
| $\rho_{pept\_a}$ | $\rho_{pept\_c}$ | 0.596 | 0.219 | 0.214 | 0.360 | 0.488 | 0.662 | 0.740 | 0.428 |
| $\rho_{pept\_b}$ | $\rho_{pept\_c}$ | 0.660 | 0.273 | 0.210 | 0.295 | 0.337 | 0.349 | 0.553 | 0.553 |

important synthesis details that the standard CC does not fully consider. Moreover, in some cases they explain an apparent disagreement between CC and visual map analysis.

With regard to an appropriate visual comparison of syntheses, we suggest rank-scaling them first and selecting the same cutoff value for the visualization of each. Alternatively, the syntheses can be taken on their initial scales (for example in $\sigma$) with the cutoff levels selected from equalization of the corresponding rank values as described in (28) and (29) in Appendix *A*.

## 4. Examples, applications and results

### 4.1. Peptide model data

We first apply the new metrics to the syntheses $\rho_{pept\_a}(\mathbf{n})$, $\rho_{pept\_b}(\mathbf{n})$ and $\rho_{pept\_c}(\mathbf{n})$ defined in §2 for a simulated peptide crystal. For the very sharp electron-density distribution $\rho_{pept\_a}(\mathbf{n})$ corresponding to a crystal with very few atoms, the rank scale is lower than that for the macromolecular syntheses at usual resolutions of 1–3 Å. In particular, for $\rho_{pept\_a}(\mathbf{n})$ the value $q = 0.80$ corresponds to a zero cutoff level in $\sigma$, the value $q = 0.95$ corresponds to $0.6\sigma$ and $q = 0.99$ corresponds to $2.2\sigma$. (Fig. 3 reminds us that for typical macromolecular syntheses the value $0\sigma$ corresponds to the range $q = 0.40$–$0.60$, the value $1\sigma$ corresponds to the range $q = 0.85$–$0.90$ and $2\sigma$ to the range $q = 0.90$–$0.95$.)

For the exact electron-density distribution $\rho_{pept\_a}(\mathbf{n})$ and the corresponding synthesis $\rho_{pept\_b}(\mathbf{n})$ at a resolution of 0.5 Å, the rank correlation coefficient is lower than the standard map correlation coefficient (Table 1). This means that for most cutoff levels the masks in the 0.5 Å resolution synthesis differ significantly from those in the exact electron density. Figs. 1(*d*) and 1(*e*) provide an example. The coefficient $CC_{90}(\rho_{pept\_a}, \rho_{pept\_b})$ is above 0.80, indicating that the peaks (their position and shape) around atomic positions are more or less conserved.

Both the CC and $CC_r$ correlation coefficients for $\rho_{pept\_a}(\mathbf{n})$ with $\rho_{pept\_c}(\mathbf{n})$ are lower than for the comparison of $\rho_{pept\_a}(\mathbf{n})$ with $\rho_{pept\_b}(\mathbf{n})$; this is owing to the lower resolution of $\rho_{pept\_c}(\mathbf{n})$ and the presence of an additional atom in the crystal. Neither of these values indicates similarity of the contour maps showing peaks above the level for the water molecule (Figs. 1*a* and 1*c*), while the correlation coefficient $CC_{95}$ does.

The Supporting Information (§S2) contains another example built on the basis of this peptide model; this example is more mathematical and illustrates comparison of grid functions by different correlation coefficients in a more transparent way. These results confirm that the new metrics describe the information contained in the crystallographic contour maps much better than the traditional metrics.

### 4.2. Incomplete low-resolution data sets

**4.2.1. Explaining an apparent contradiction between low correlation coefficients and similar contour maps**. A model $F_{calc}\exp(i\varphi_{calc})$ Fourier synthesis (referred to as $\rho_{incomplete}$) computed for PDB entry 1nh2 by (2) using reflection indices from the deposited data set (Bleichenbacher *et al.*, 2003; highest resolution 1.9 Å; data completeness 95%) shows part of the structure very poorly (Fig. 2*a*). Fig. 2(*b*) shows a model Fourier synthesis $\rho_{complete}$ calculated with the same coefficients

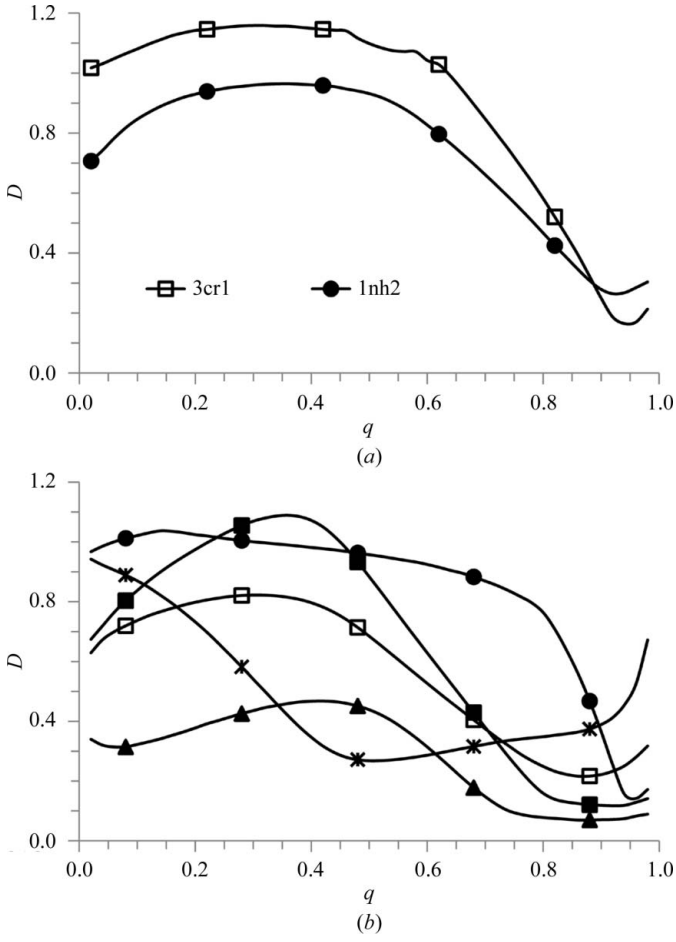


**Figure 4**
Discrepancy function $D(q)$ comparing the Fourier contour maps obtained with complete and incomplete data sets. (*a*) Comparison of the syntheses $\rho_{incomplete}$ calculated for the set of reflections as deposited in the PDB with the syntheses $\rho_{complete}$ obtained with the complete data set of the respective resolution ($d_{high} = 1.90$ Å for 1nh2 and $d_{high} = 2.25$ Å for 3cr1). (*b*) Comparison of $\rho_{complete}$ (complete data set at a resolution from $d_{high}$ to infinity) with $\rho_{incomplete}$ calculated with a data set in the resolution interval $d_{high}$ to 10 Å. The curves are shown for $d_{high} = 2$ Å (solid squares, 1zud; solid circles, 1q09; solid triangles, 1ous) and for $d_{high} = 4$ Å (open squares, 1zud). The curve marked by stars is for comparison of the two $\rho_{complete}$ maps for 1zud, one calculated with the complete data set at a resolution of 2 Å and the other with the complete data set at a resolution of 4 Å.

**Table 2**
Comparison of the Fourier syntheses for selected PDB entries.

All syntheses were obtained with the calculated structure factors ($F_{calc}$, $\varphi_{calc}$). Correlation coefficients between $\rho_a$ and $\rho_b$ are defined in the text.

| PDB code | $\rho_a$ resolution (Å) | $\rho_b$ resolution (Å) | CC | $CC_r$ | $CC_{50}$ | $CC_{70}$ | $CC_{80}$ | $CC_{90}$ | $CC_{95}$ | $CC_{99}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1q09 | 2–∞ | 2–10 | 0.714 | 0.059 | 0.080 | 0.242 | 0.471 | 0.811 | 0.850 | 0.669 |
| | 4–∞ | 4–10 | 0.605 | 0.495 | 0.353 | 0.388 | 0.510 | 0.680 | 0.591 | 0.238 |
| | 4–∞ | 2–∞ | 0.876 | 0.441 | 0.560 | 0.825 | 0.847 | 0.629 | 0.438 | −0.070 |
| 1zud | 2–∞ | 2–10 | 0.855 | 0.443 | 0.666 | 0.864 | 0.908 | 0.899 | 0.875 | 0.845 |
| | 4–∞ | 4–10 | 0.764 | 0.622 | 0.670 | 0.764 | 0.767 | 0.683 | 0.590 | 0.459 |
| | 4–∞ | 2–∞ | 0.797 | 0.779 | 0.685 | 0.555 | 0.468 | 0.260 | 0.002 | −0.402 |
| 1ous | 2–∞ | 2–10 | 0.961 | 0.860 | 0.868 | 0.958 | 0.962 | 0.954 | 0.938 | 0.908 |
| | 4–∞ | 4–10 | 0.888 | 0.832 | 0.808 | 0.846 | 0.857 | 0.820 | 0.745 | 0.588 |
| | 4–∞ | 2–∞ | 0.596 | 0.492 | 0.419 | 0.331 | 0.239 | 0.005 | −0.250 | −0.528 |

using all theoretically possible reflections up to 1.9 Å resolution. The correlation coefficient CC calculated using (4) between the two syntheses is 0.70. Since both syntheses were calculated with the model data, the only source of difference is the missing reflections, essentially the lowest resolution reflections (there are 300 reflections missing from 408 with resolution below 10 Å; all 59 reflections with resolution below 20 Å are missing).

A similar comparison of $\rho_{incomplete}$ with $\rho_{complete}$ for another test case (PDB entry 3cr1; MacElrevey *et al.*, 2008; highest resolution 2.25 Å; data completeness 98%) yields an even lower map correlation coefficient CC = 0.64, which one would expect to be reflected by a larger difference between the two maps. This low correlation coefficient is owing to missing only 2% of the reflections (there are 116 reflections missed out of 251 collected at a resolution below 10 Å and 32 reflections out of 42 at a resolution below 20 Å). However, the contour map obtained with the incomplete data set is perfectly interpretable for the whole molecule and is very similar to the map calculated with the complete set of reflections (compare Fig. 2*d* with Fig. 2*c*). This illustrates that the map correlation coefficient is not necessarily a good predictor of the visual similarity of maps either from its value or when comparing different pairs of maps.

The rank correlation coefficient $CC_r$ (18) is 0.30 for 1nh2 and just 0.01 for 3cr1 and is even lower than the values of the standard map correlation coefficient. It shows that in this case of missing low-resolution data most of the masks are severely changed compared with the corresponding masks in $\rho_{complete}$ (see also Urzhumtsev, 1991; Urzhumtseva & Urzhumtsev, 2011).

The peak correlation coefficient considers only the part of the map in the quantile rank greater than 0.90 and gives different information. Its value is 0.67 for 1nh2 and 0.83 for 3cr1 and shows that the peaks are conserved much better for 3cr1, agreeing with the visual analysis. This relationship is not shown by either the standard map correlation coefficient or the rank correlation coefficient. Fig. 4(*a*) expands on this calculation of $CC_{90}$ by showing the discrepancy function $D(q)$ for these two comparisons. It can be seen that for most rank values $q$ the contours are quite different in both cases, $D(q) \simeq 1$, that for high $q$ values such as 0.90 they are equally similar and for very high values such as $q \simeq 0.95$ they are more similar for 3cr1.

**4.2.2. Effect of low-resolution incompleteness on crystals with various solvent contents**. The examples in §4.2.1 illustrate the effect of low-resolution data incompleteness. It seemed possible that the strength of this effect might depend on the fraction of bulk solvent in the crystal. We made a comparative analysis considering three cases of bulk-solvent content: near the very common value of 50% (PDB entry 1zud; Lehmann *et al.*, 2006; solvent content 0.47), very high (PDB entry 1q09; Changela *et al.*, 2003; solvent content 0.84) and very low (PDB entry 1ous; Loris *et al.*, 2003; solvent content 0.24).

For each of these structures, we calculated a complete set of structure factors $F_{calc}\exp(i\varphi_{calc})$ from the atomic model at a resolution of 2 Å. We call the Fourier synthesis calculated with these structure factors $\rho_{2-\infty}(\mathbf{n}) = \rho_{complete}(\mathbf{n})$. We also calculated another Fourier synthesis $F_{calc}\exp(i\varphi_{calc})$ in which all of the structure factors at a resolution outside the range 2–10 Å were excluded. We call this synthesis omitting low-resolution data beyond 10 Å $\rho_{2-10}(\mathbf{n}) = \rho_{incomplete}(\mathbf{n})$.

Both the conventional correlation coefficient CC and the rank correlation coefficient $CC_r$ comparing $\rho_{2-\infty}(\mathbf{n})$ and $\rho_{2-10}(\mathbf{n})$ decrease with increasing volume of the bulk-solvent region in these cases (Table 2). Note that for 1ous, with an extremely low bulk-solvent content, all of the maps are well conserved.

The variation in $CC_r$ is more significant; in particular, its value of close to zero for 1q09 means that for these data most of the masks changed, information which is difficult to extract from the CC value of above 0.7. At the same time, the peaks are well conserved for all three structures; Fig. 5 gives an example for 1zud. The larger the bulk-solvent content, the higher the quantile rank corresponding to the highest value of the peak correlation coefficient (Table 2; Fig. 4*b*).

The situation is quantitatively similar when we compare the corresponding maps $\rho_{4-\infty}(\mathbf{n}) = \rho_{complete}(\mathbf{n})$ and $\rho_{4-10}(\mathbf{n}) = \rho_{incomplete}(\mathbf{n})$ calculated with data at lower resolution, in the ranges from 4 Å to infinity and from 4 to 10 Å, respectively (Table 2).

### 4.3. Effect of data-resolution cutoff

Intuitively, it is clear that excluding high-resolution data changes the maps in a different way than excluding low-resolution data. It is easy to illustrate this using the new metrics.

To do so, for each of the three structures described in §4.2 we calculated the $F_{calc}\exp(i\varphi_{calc})$ syntheses $\rho_{2-\infty}(\mathbf{n})$ and $_{4-\infty}(\mathbf{n})$ with the complete data sets at resolutions of 2 and 4 Å, respectively, and compared them. The map correlation coef-

ficient values $CC(\rho_{2-\infty}, \rho_{4-\infty})$ are relatively high; for example for 1q09 this coefficient is as high as 0.88. This number shows some difference in the maps at such high- and low-resolution cutoffs; however, one might intuitively expect a much larger difference. Indeed, the rank correlation coefficient $CC_r(\rho_{2-\infty}, \rho_{4-\infty})$ is much lower for 1q09, being equal to 0.44 and showing that the maps are substantially different.

As expected, the peak correlation coefficients for high rank values $q$ are low (see, for example, $CC_{95}$ and $CC_{99}$) since the peaks are merged in the 4 Å resolution maps in comparison with the 2 Å resolution maps. The close-to-zero values of these coefficients are more intuitive than the value of the map correlation coefficient $CC(\rho_{2-\infty}, \rho_{4-\infty})$ given above.

At the same time, some peak correlation coefficients are relatively high, *e.g.* $CC_{80}(\rho_{2-\infty}, \rho_{4-\infty}) = 0.85$ for 1q09. The corresponding rank value corresponds well to that defining the molecular region (see also Fig. 5) and shows that the molecular masks are less affected by excluding the high-resolution data. For the 1ous data, the molecule occupies practically the whole unit cell (and simply the whole unit cell if structural waters are included), and all peak correlation coefficients for it are low, showing changes in the maps at all cutoff levels.

Thus, using $CC_r$ and the rank correlation coefficients may illustrate features that are difficult to see when referring only to the standard map correlation coefficient CC (4).

### 4.4. Effect of excluding reflections for cross-validation

§4.2 shows that the loss of a relatively small number of low-resolution reflections (as few as 2%) can result in significant changes in the Fourier contour maps. On the other hand, the test data set (Brünger, 1992), typically containing 5–10% of the total number of reflections, is purposely excluded from all calculations; this should be the case for all steps including,

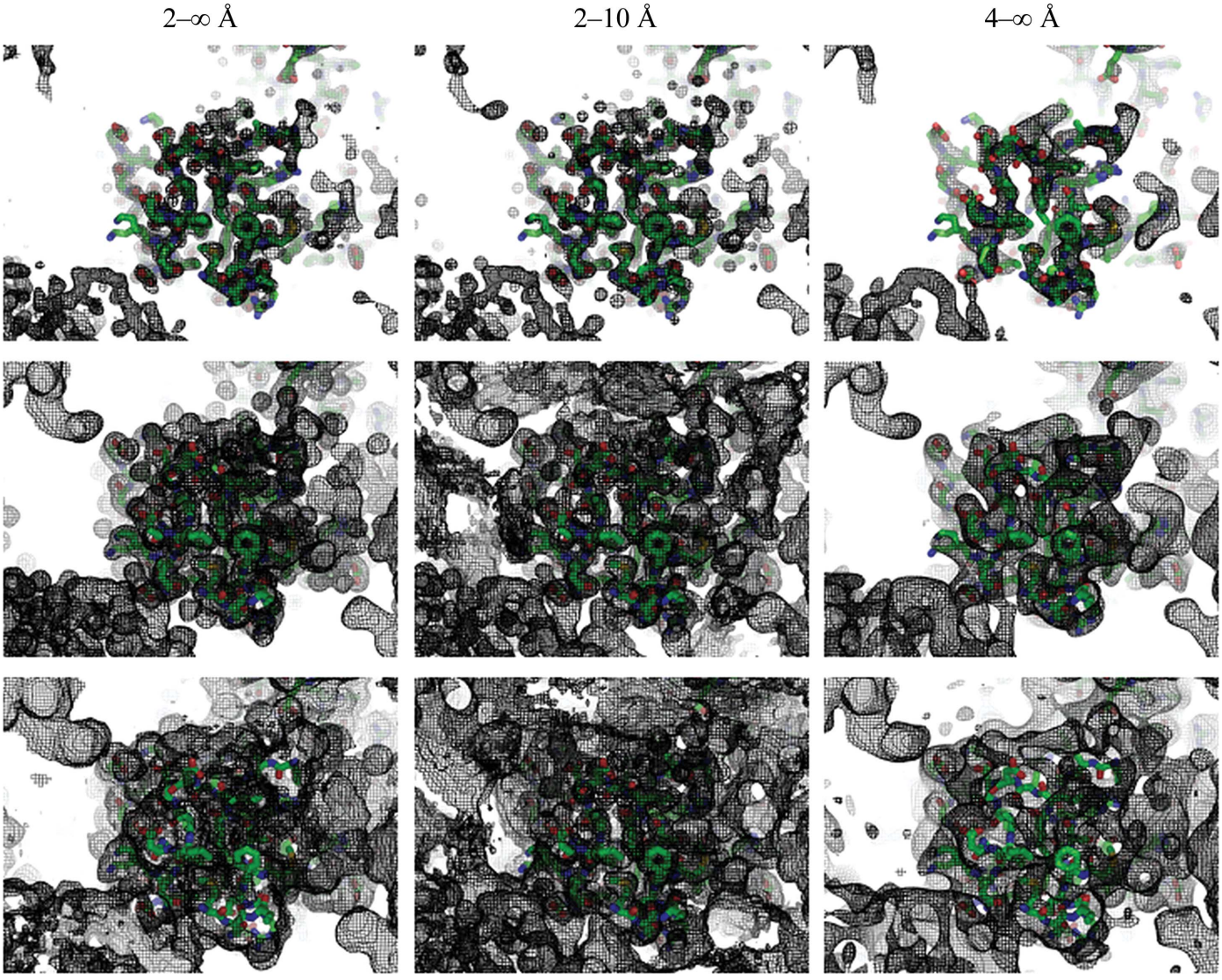


**Figure 5**
Fourier contour maps for 1zud. All syntheses are calculated with the model structure factors at the resolution cutoff indicated above each column and are shown at different rank levels (top, 0.9; middle, 0.7; bottom, 0.5). The solvent content is 0.47. Note the conservation of peaks and the loss of the molecular envelope when removing low-resolution data and the conservation of the envelope with decreasing resolution.

formally speaking, the calculation of contour maps (although the latter is not always the case in practice). These data are used for validation (Brünger, 1992) and to estimate statistical parameters (Lunin & Skovoroda, 1995; Pannu & Read, 1996; Murshudov *et al.*, 1997). In general, the reflections for the test set are chosen randomly and uniformly across reciprocal space.

There is an old and frequently asked question whether excluding such reflections noticeably distorts the Fourier contour maps. We do not analyze this question in detail here, but simply illustrate the effects for a typical protein structure under typical conditions. To do so, we used the IF2 structure that we recently solved (Simonetti *et al.*, 2013; PDB entry 4b3x). The corresponding crystals belonged to space group $P2_12_12_1$, with unit-cell parameters $a$ = 45.42, $b$ = 61.46, $c$ = 162.40 Å. The experimental data set is complete to 2 Å resolution (with only two low-resolution reflections missing); bulk solvent occupies approximately 50% of the unit cell. The $R$ and $R_{\rm free}$ values calculated by *PHENIX* (Adams *et al.*, 2010) are less than 0.18 and 0.22, respectively, showing that the structure factors $F_{\rm model}\exp(i\varphi_{\rm model})$ calculated from the atomic model including the correction from the bulk solvent [Jiang & Brünger (1994), with the improvements described by Afonine *et al.* (2013)], reproduce the experimental data well. Thus, we used the phase values $\varphi_{\rm model}$ as the best possible approximation to the unknown values to be associated with the experimental structure-factor amplitudes $F_{\rm obs}$.

We calculated a series of Fourier syntheses at a resolution of 2 Å with coefficients $F_{\rm obs}\exp(i\varphi_{\rm model})$, with the fraction of randomly excluded reflection ranging between 5 and 10%, as is routinely undertaken for test-set reflections. Each of these syntheses was compared with a synthesis calculated with the complete data set. The correlation coefficient CC between them remained high, *i.e.* above 0.90, even when the test set contained up to 20% of the data. However, the peak correlation coefficients $CC_{50}$–$CC_{80}$ indicated non-negligible map changes when 10% of the data were excluded. The maps showed significant noise at the rank value $q$ = 0.80 (roughly 0.4$\sigma$ for this synthesis) and incorrect density for a few weakly defined side chains. We note that the molecule occupies approximately half of the unit cell: $q$ = 0.50. The differences resulting from the exclusion of 10% of reflections are more significant than the differences owing to experimental errors in amplitudes, as can be seen from comparison with the maps calculated with coefficients $F_{\rm model}\exp(i\varphi_{\rm model})$ (Table 3). Overall, maps obtained with the model data $F_{\rm model}\exp(i\varphi_{\rm model})$ illustrated a behaviour similar to that for $F_{\rm obs}\exp(i\varphi_{\rm model})$ maps.

**Table 3**
Influence of excluded test data sets.

Fourier syntheses at the resolution $d_{\rm high}$ = 2 Å were calculated for the IF2 structure (Simonetti *et al.*, 2013) using $F_{\rm obs}$ or $F_{\rm model}$ amplitudes and phases $\varphi_{\rm model}$. Correlation coefficients between $\rho_a$ and $\rho_b$ are defined in the text.

| Type of amplitudes | $\rho_a$, data excluded (%) | $\rho_b$, data excluded (%) | CC | $CC_r$ | $CC_{50}$ | $CC_{70}$ | $CC_{80}$ | $CC_{90}$ | $CC_{95}$ | $CC_{99}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $F_{\rm obs}$ | 0 | 5 | 0.976 | 0.900 | 0.783 | 0.805 | 0.938 | 0.966 | 0.957 | 0.905 |
| | | 10 | 0.951 | 0.842 | 0.694 | 0.786 | 0.887 | 0.936 | 0.920 | 0.834 |
| | | 20 | 0.899 | 0.753 | 0.591 | 0.684 | 0.792 | 0.869 | 0.841 | 0.705 |
| $F_{\rm model}$ | 0 | 5 | 0.974 | 0.850 | 0.687 | 0.846 | 0.952 | 0.963 | 0.954 | 0.895 |
| | | 10 | 0.950 | 0.797 | 0.627 | 0.788 | 0.905 | 0.935 | 0.918 | 0.824 |
| | | 20 | 0.900 | 0.715 | 0.547 | 0.691 | 0.813 | 0.873 | 0.840 | 0.685 |

**Table 4**
Influence of amplitudes and bulk-solvent modelling.

The Fourier syntheses were calculated with the complete data sets at resolution $d_{\rm high}$ for the IF2 structure (Simonetti *et al.*, 2013). Correlation coefficients between $\rho_a$ and $\rho_b$ are defined in the text.

| $d_{\rm high}$ (Å) | $\rho_a$, coefficients | $\rho_b$, coefficients | CC | $CC_r$ | $CC_{50}$ | $CC_{70}$ | $CC_{80}$ | $CC_{90}$ | $CC_{95}$ | $CC_{99}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | $F_{\rm obs}$, $\varphi_{\rm model}$ | $F_{\rm model}$, $\varphi_{\rm model}$ | 0.970 | 0.834 | 0.671 | 0.843 | 0.940 | 0.944 | 0.926 | 0.819 |
| 2 | $F_{\rm calc}$, $\varphi_{\rm calc}$ | $F_{\rm model}$, $\varphi_{\rm model}$ | 0.894 | 0.623 | 0.712 | 0.905 | 0.940 | 0.963 | 0.961 | 0.953 |
| 3 | $F_{\rm calc}$, $\varphi_{\rm calc}$ | $F_{\rm model}$, $\varphi_{\rm model}$ | 0.881 | 0.684 | 0.733 | 0.915 | 0.949 | 0.941 | 0.925 | 0.864 |

Summarizing, we suggest that carrying out an analysis of the rank and peak correlation coefficients could be used as a routine tool for identifying a suitable fraction of reflections for a test set in Fourier syntheses even when this set has been already assigned. A synthesis may be calculated with the working set of reflections and with the full available data set, and if the rank or peak correlation coefficients between these maps are low (a more systematic analysis is probably required to define appropriate critical values), the test data set might be reduced for further calculations by reassigning, also randomly and uniformly, some reflections back to the working set. As this example shows, the usual correlation coefficient alone may be not sufficiently informative.

### 4.5. Bulk-solvent contribution

It is largely accepted that using a bulk-solvent correction is vital in order to properly include low-resolution data into the structure-solution process (see, for example, Phillips, 1980; Fenn *et al.*, 2010; Afonine *et al.*, 2013 and references therein). However, the influence of the bulk-solvent correction on Fourier syntheses has been less discussed.

To analyze the direct effect of the bulk-solvent contribution on the Fourier synthesis, complementary to the synthesis with $\{F_{\rm model}\exp(i\varphi_{\rm model})\}$ for the IF2 model (§4.4), we calculated another synthesis with the structure factors $\{F_{\rm calc}\exp(i\varphi_{\rm calc})\}$ without a bulk-solvent correction. The data sets were complete at the resolution of 2 Å. As mentioned above, the first data set, including the bulk solvent, reproduces the experimental data quite well.

The correlation coefficient CC between the two syntheses, equal to 0.89, indicates their high similarity. However, the rank coefficient $CC_r$ of 0.62 shows that in fact the changes in the

map owing to unmodelled bulk solvent are not negligible. This means that ignoring a bulk-solvent correction when modelling the 'experimental syntheses' may result in maps that differ from the correct maps and therefore may lead to wrong or unjustified conclusions. In particular, such data are not recommended for analysis of molecular envelopes since they may be mostly affected by this improper modelling (Table 4). At the same time, such simulated syntheses can be successfully used when studying only the structural details since $CC_{80}$–$CC_{95}$ indicate very high similarity of the peaks.

Comparison of the corresponding syntheses calculated at a resolution of 3 Å gives values comparable with those for the 2 Å resolution syntheses. However, the peak correlation coefficients for the rank $q \geq 0.9$ are lower. For example, the coefficient $CC_{99}$ corresponding roughly to the $3\sigma$ cutoff level decreases from 0.95 at 2 Å to 0.86 at 3 Å. This indicates that at lower resolution limits the unmodelled bulk-solvent contribution may distort not only the molecular envelopes but also the peaks of the syntheses.

## 5. Discussion

The several examples presented in this work show that the traditional map correlation coefficient CC does not always correspond well to the similarity of or the difference in two Fourier syntheses based on visual examination. Approaches are presented to address this problem. They are based on the concept of a rank scaling of the syntheses. With such a scaling, regions selected with the same cutoff level contain the same number of grid nodes and the number of grid nodes in common is a useful measure of the similarity of the maps at that cutoff level.

The rank correlation coefficient $CC_r$ is calculated as a correlation of the rank-scaled syntheses instead of the initial values $\rho(\mathbf{n})$, for example those in $\sigma$. Both CC and $CC_r$ are equal to 1 when the values of the two syntheses are related by a linear transformation. However, in contrast to CC, $CC_r$ is equal to 1 also when the values of the syntheses are related by a nonlinear monotonic transformation; here, the maps are exactly the same but correspond to different cutoff levels on the original scales.

To accompany traditional visual analysis, we suggest using the peak correlation coefficients, in particular $CC_{90}$, as a rule of thumb, adjusting the peak level to particular situations and problems. To compare molecular masks or peaks in the low-resolution maps, the correlation coefficient $CC_{50}$ may be more appropriate. The discrepancy function $D(q; \rho_a, \rho_b)$ compares the selected regions (masks) by counting the number of grid nodes in complementary regions, regardless of the exact values of the syntheses in these nodes.

The computational tools described here may be applied to answer additional questions to those that we have illustrated. The new coefficients may be calculated not in the whole unit cell but locally in a given region. With the peak correlation coefficient, one may compare syntheses previously difficult to compare numerically such as the usual $\sigma_A$ synthesis and a difference synthesis. These tools may be used, in the case of comparing several maps, to select the one for which the corresponding contour maps correspond better to a control map. Naturally, the choice of the map for comparison is important and should be considered for each particular project.

The developed metrics can be also applied to compare maps corresponding to different crystals or to noncrystallographic objects, for example electron microscopy reconstructed images. The only requirement is that the compared parts of the images are of the same size and the maps are calculated on the same grid.

The tools discussed in this manuscript, namely the discrepancy function $D(q; \rho_a, \rho_b)$, the rank correlation coefficient $CC_r(\rho_a, \rho_b)$ and the peak correlation coefficient $CC_{<q\text{peak}>}(\rho_a, \rho_b)$, are implemented in *PHENIX* (Adams *et al.*, 2010) and are also available as an independent program from AU.

## APPENDIX *A*
## Rank-scaled synthesis and corresponding statistical moments

Firstly, for a synthesis calculated on an arbitrary scale on a grid composed of $N_{\text{grid}}$ nodes $\mathbf{n}$, one computes the frequency of its values, as was introduced into crystallography by Lunin (1988) and Main (1990*a*,*b*). To do so, the interval $(\rho_{\min}, \rho_{\max}) = [\min \rho(\mathbf{n}), \max \rho(\mathbf{n})]$ is divided into $J$ nonintersecting subintervals called bins: $(\rho_0 = \rho_{\min}, \rho_1)$, $(\rho_1, \rho_2)$, $\ldots$, $(\rho_{J-1}, \rho_J = \rho_{\max})$. For each grid node $\mathbf{n}$, we identify the interval $j$ to which the corresponding value $\rho(\mathbf{n})$ belongs to,

$$\rho_{j-1} \leq \rho(\mathbf{n}) < \rho_j \tag{24}$$

and add a unit to the counter $n_j$ of this bin. The frequencies of the synthesis values are then calculated as

$$\nu_j = \frac{n_j}{N_{\text{grid}}}, \quad j = 1, 2, \ldots, J, \tag{25}$$

giving

$$\sum_{j=1}^{J} \nu_j = 1. \tag{26}$$

The quantile ranks $q_j$ corresponding to the bin borders $\mu = \rho_j$ are the numbers $N(\rho_j)$ of grid nodes $\mathbf{n}$ with the synthesis value below the threshold, $\rho(\mathbf{n}) < \rho_j$, normalized as

$$q_j = \frac{N(\rho_j)}{N_{\text{grid}}} = \sum_{i \leq j} \nu_i. \tag{27}$$

For an intermediate value $\rho_{j-1} < \mu < \rho_j$, its quantile rank $q$ may be calculated by a linear interpolation

$$q = q_{j-1} + (q_j - q_{j-1}) \frac{\mu - \rho_{j-1}}{\rho_j - \rho_{j-1}}, \tag{28}$$

making it a strictly increasing function of the initial synthesis values. Inversely, for a given rank $q_{j-1} < q < q_j$ the corresponding initial synthesis value is recovered as

$$\rho = \rho_{j-1} + (\rho_j - \rho_{j-1}) \frac{q - q_{j-1}}{q_j - q_{j-1}}. \tag{29}$$

The value complementary to $q(\mu;\ \rho)$ gives the fractional volume $V_{\rm f}$ of the unit cell selected by the corresponding cutoff level

$$V_{\rm f} = 1 - q(\mu;\ \rho) = \frac{N_{\rm grid} - N(\mu)}{N_{\rm grid}} \simeq \frac{\text{Volume}\{xyz : \rho(xyz) > \mu\}}{\text{Volume}\{\text{unit cell}\}}. \quad (30)$$

Scaling of crystallographic Fourier syntheses in fractional volume has been described previously, for example by Vernoslova & Lunin (1993).

Let $\mathrm{Q}(\mathbf{n})$ be a rank-scaled synthesis where each value $\rho(\mathbf{n})$ is substituted by the corresponding quantile rank using (28). We split the nodes into $M$ equal groups defined by the values of the synthesis, with no relation to their position in the cell. The first $N_{\rm grid}$ points correspond to the lowest values of the synthesis; the corresponding rank values are $0 < Q(\mathbf{n}) < 1/M$; the next $N_{\rm grid}/M$ points correspond to slightly higher values with $1/M < Q(\mathbf{n}) < 2/M$ *etc*. Then, for large enough $M$,

$$\begin{aligned}\sum_{\mathbf{n}} Q(\mathbf{n}) &\simeq \left(\frac{N_{\rm grid}}{M}\frac{1}{M} + \frac{N_{\rm grid}}{M}\frac{2}{M} + \ldots + \frac{N_{\rm grid}}{M}\frac{M}{M}\right) \\ &= \frac{N_{\rm grid}}{M}\frac{1}{M}(1 + 2 + \ldots + M) \\ &= \frac{N_{\rm grid}}{M}\frac{1}{M}\frac{M(M+1)}{2} \simeq \frac{1}{2}N_{\rm grid}. \end{aligned} \quad (31)$$

Similarly,

$$\begin{aligned}\sum_{\mathbf{n}} Q^2(\mathbf{n}) &= \frac{N_{\rm grid}}{M}\left(\frac{1}{M}\right)^2 + \frac{N_{\rm grid}}{M}\left(\frac{2}{M}\right)^2 + \ldots + \frac{N_{grid}}{M}\left(\frac{M}{M}\right)^2 \\ &= \frac{N_{\rm grid}}{M}\frac{1}{M^2}(1^2 + 2^2 + \ldots + M^2) \\ &= \frac{N_{\rm grid}}{M}\frac{1}{M^2}\frac{M(M+1)(2M+1)}{6} \simeq \frac{1}{3}N_{\rm grid}. \end{aligned} \quad (32)$$

The contour maps used in this work and shown in Figs. 1, 2 and 5 and in the Supporting Information were produced using *PyMOL* (DeLano, 2002). PVA, TCT and PDA thank the NIH (grant GM063210) and the PHENIX Industrial Consortium for support of the PHENIX project. This work was supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231 (PVA, TCT and PDA) and by Russian Foundation for Basic Research grant 13-04-00118-a (VYL). AU thanks the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INSB-05-01 and Instruct as part of the European Strategy Forum on Research Infrastructures (ESFRI). We thank the referees for their very fruitful and constructive comments.

## References


Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D. & Urzhumtsev, A. (2013). *Acta Cryst.* D**69**, 625–634.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Bhat, T. N. (1988). *J. Appl. Cryst.* **21**, 279–281.
Bleichenbacher, M., Tan, S. & Richmond, T. J. (2003). *J. Mol. Biol.* **332**, 783–793.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Brändén, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Bruckner, S. & Möller, T. (2010). *Comput. Graph. Forum*, **29**, 773–782.
Burla, M. C., Caliandro, R., Giacovazzo, C. & Polidori, G. (2010). *Acta Cryst.* A**66**, 347–361.
Changela, A., Chen, K., Xue, Y., Holschen, J., Outten, C. E., O'Halloran, T. V. & Mondragon, A. (2003). *Science*, **301**, 1383–1387.
DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Ewald, P. P. (1913). *Phys. Z.* **14**, 465–472.
Fenn, T. D., Schnieders, M. J. & Brunger, A. T. (2010). *Acta Cryst.* D**66**, 1024–1031.
Gore, S., Velankar, S. & Kleywegt, G. J. (2012). *Acta Cryst.* D**68**, 478–483.
Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.
Jiang, J.-S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.
Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Lang, P. T., Holton, J. M., Fraser, J. S. & Alber, T. (2014). *Proc. Natl Acad. Sci. USA*, **111**, 337–342.
Lehmann, C., Begley, T. P. & Ealick, S. E. (2006). *Biochemistry*, **45**, 11–19.
Lehmann, E. L. & D'Abrera, H. J. M. (1998). *Nonparametrics: Statistical Methods Based on Ranks*. Englewood Cliffs: Prentice–Hall.
Loris, R., Tielker, D., Jaeger, K. E. & Wyns, L. (2003). *J. Mol. Biol.* **331**, 861–870.
Lunin, V. Y. (1988). *Acta Cryst.* A**44**, 144–150.
Lunin, V. Y. (1989). *Acta Cryst.* A**45**, 501–505.
Lunin, V. Y. (1993). *Acta Cryst.* D**49**, 90–99.
Lunin, V. Y., Lunina, N. L. & Urzhumtsev, A. G. (2000). *Acta Cryst.* A**56**, 375–382.
Lunin, V. Y. & Skovoroda, T. P. (1995). *Acta Cryst.* A**51**, 880–887.
Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.
Luzzati, V. (1953). *Acta Cryst.* **6**, 142–152.
MacElrevey, C., Salter, J. D., Krucinska, J. & Wedekind, J. E. (2008). *RNA*, **14**, 1600–1616.
Main, P. (1979). *Acta Cryst.* A**35**, 779–785.
Main, P. (1990*a*). *Acta Cryst.* A**46**, 372–377.
Main, P. (1990*b*). *Acta Cryst.* A**46**, 507–509.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Phillips, S. E. V. (1980). *J. Mol. Biol.* **142**, 531–554.
Pozharski, E. (2010). *Acta Cryst.* D**66**, 970–978.
Pratt, W. K. (1978). *Digital Image Processing*. New York: Wiley.
Ramachandran, G. N. & Raman, S. (1959). *Acta Cryst.* **12**, 957–964.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Rupp, B. (2006). *Nature (London)*, **444**, 817.
Schotte, F., Lim, M., Jackson, T. A., Smirnov, A. V., Soman, J., Olson, J. S., Phillips, G. N. Jr, Wulff, M. & Anfinrud, P. A. (2003). *Science*, **300**, 1944–1947.
Simonetti, A., Marzi, S., Fabbretti, A., Hazemann, I., Jenner, L., Urzhumtsev, A., Gualerzi, C. O. & Klaholz, B. P. (2013). *Acta Cryst.* D**69**, 925–933.
Spearman, C. (1904). *Am. J. Psychol.* **15**, 72–101.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.

Tickle, I. J. (2012). *Acta Cryst.* D**68**, 454–467.
Urzhumtsev, A. G. (1991). *Acta Cryst.* A**47**, 794–801.
Urzhumtsev, A. G., Lunin, V. Y. & Luzyanina, T. B. (1989). *Acta Cryst.* A**45**, 34–39.
Urzhumtseva, L. & Urzhumtsev, A. (2011). *J. Appl. Cryst.* **44**, 865–872.
Vernoslova, E. A. & Lunin, V. Y. (1993). *J. Appl. Cryst.* **26**, 291–294.
Vijayan, M. (1980). *Acta Cryst.* A**36**, 295–298.