Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the interpretation and use of $\langle |E|^2 \rangle(d^*)$ profiles

Richard J. Morris,† Eric Blanc† and Gérard Bricogne

Global Phasing Ltd, Sheraton House, Castle Park, Cambridge CB3 0AX, England

† Present address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England.

Received 26 September 2003
Accepted 5 November 2003

Profiles of squared normalized structure factors, $\langle |E|^2 \rangle(d^*)$, have been computed for a large number of proteins and nucleic acids. These are interpreted in terms of their underlying structural features. It is also shown that the 'solvent dip' at around 6.3 Å resolution is to a large extent a protein secondary-structure effect that is enhanced by the water structure. A hierarchical classification of protein structures based on their $\langle |E|^2 \rangle(d^*)$ profiles is briefly outlined, together with the use of $\langle |E|^2 \rangle(d^*)$ profiles as an improvement over Wilson absolute scale estimation and as a novel solvent-modelling method.

1. Introduction

The concept of randomly, uniformly and independently distributed atoms, although clearly unable to capture the structural regularities of chemical bonding and macromolecular architecture, continues to play a central rôle in virtually all statistical steps in crystallographic structure solution and refinement.

Deviations from the hypothesis of uniform random distributions, while that of independent atoms is retained, were investigated by Luzzati (1955) and their exploitation in structure solution was undertaken by Bricogne (1984, 1988, 1991, 1993, 1997*a*). Deviations from the hypothesis of independence, on the other hand, have been studied by Harker (1953), Main (1976) and others. In a previous paper (Morris & Bricogne, 2003), we investigated them by multipole-expansion techniques proposed in the context of the 'micromolecular-replacement' approach to exploiting stereochemistry in statistical direct methods (Bricogne, 1994, 1995, 1997*b*). This study brought to light a connection with Sheldrick's '1.2 Å rule' (Sheldrick, 1990) in traditional direct methods.

In this article, we first offer a brief survey of topics relevant to normalized structure factors, then investigate the full resolution range of $\langle |E|^2 \rangle(d^*)$ curves and analyse the relationships between the features observed in these curves and the known structural features of both proteins and nucleic acids. We finally examine the question of correcting $\langle |E|^2 \rangle(d^*)$ profiles for bulk-solvent effects and assess their usefulness in the determination of absolute scale and in the normalization of observed structure-factor amplitudes during structure determination and in the refinement of partial structure models.

2. Normalized structure factors

The normalization process was introduced by Hauptman & Karle (1953) as a means of turning the X-ray diffraction

intensities[1] into dimensionless quantities which characterize the internal atomic arrangements within the crystal, as independently as possible from the chemical identity and disorder characteristics of its constituent atoms. The resulting normalized structure-factor amplitudes turn out to be the natural quantities by which to assess the strength of phase relationships in direct methods.

More specifically, each observed structure-factor amplitude $|F_{\rm obs}(\mathbf{h})|$ is normalized to $|E(\mathbf{h})|$ by division by a resolution-dependent quantity such that the expectation value $\langle|E(\mathbf{h})|^2\rangle$ for any given $\mathbf{h} = (hkl)^T$ is unity,

$$\langle|E(\mathbf{h})|^2\rangle = 1.0. \tag{1}$$

This expectation value is formally achieved by setting

$$|E(\mathbf{h})|^2 = \frac{|F_{\rm obs}(\mathbf{h})|^2}{\langle|F(\mathbf{h})|^2\rangle}, \tag{2}$$

where $|F(\mathbf{h})|^2$ is the computed squared structure-factor amplitude and $\langle|F(\mathbf{h})|^2\rangle$ is its expectation value, which depends on the prior knowledge one has at hand about the structure. With no structural information, the atomic positions are considered to be uniformly and independently distributed random variables and the expectation of the squared structure-factor amplitude is equal to

$$\langle|F(\mathbf{h})|^2\rangle = \sum_{i=1}^{N}[f_i(\mathbf{h})W_i(\mathbf{h})]^2 \tag{3}$$

in which the Debye–Waller factor has been defined according to

$$W_i(\mathbf{h}) = \exp(-\mathbf{h}^T\beta_i\mathbf{h}), \tag{4}$$

with β_i as the atomic displacement factor tensor and $f_i(\mathbf{h}) = f_i(d^*_{\mathbf{h}})$ the spherically symmetric atomic scattering factor of atom i. N is the total number of atoms.

If the composition of the molecule is defined in terms of known groups, G, assumed to be independently distributed in position and orientation, then Main (1976) showed that the expectation value may be written as

$$\langle|F(\mathbf{h})|^2\rangle = \sum_{i=1}^{N}[f_i(d^*_{\mathbf{h}})W_i(\mathbf{h})]^2 + \sum_G \sum_{i_G\neq j_G} f_{i_G}(d^*_{\mathbf{h}})W_{i_G}(\mathbf{h})f_{j_G}(d^*_{\mathbf{h}})W_{j_G}(\mathbf{h})\frac{\sin(2\pi d^*|\mathbf{r}_{j_G}-\mathbf{r}_{i_G}|)}{2\pi d^*|\mathbf{r}_{j_G}-\mathbf{r}_{i_G}|} \tag{5}$$

in which the $f_{j_G}(d^*)$ are spherically averaged group scattering factors.

The more structural knowledge is correctly accounted for, the more the expectation of the squared normalized structure-factor amplitudes should approach unity.

[1] In this article, by intensities we mean *corrected intensities* (squared structure-factor amplitudes), *i.e.* corrected for Lorentz factor, polarization, absorption *etc*.

2.1. Wilson statistics and standard normalized structure factors

If the distribution of random atomic positions within the unit cell is uniform and the atoms are independent, the distribution of $F(\mathbf{h})$ can readily be shown to be Gaussian: a one-dimensional Gaussian with variance $\Sigma_c(\mathbf{h}) = \varepsilon(\mathbf{h})\sigma_2$ for centric reflections and a two-dimensional Gaussian with variance $\Sigma_a(\mathbf{h}) = \frac{1}{2}\varepsilon(\mathbf{h})\sigma_2(\mathbf{h})$ for the acentric reflections (Wilson, 1949, 1950), where $\sigma_2(\mathbf{h}) = \sum f_j^2(\mathbf{h})$ and $\varepsilon(\mathbf{h})$ is the multiplicity of reflection $\mathbf{h}$ (Stewart & Karle, 1976; Stewart *et al.*, 1977).

Standard normalized structure-factor amplitudes are defined by

$$|E(\mathbf{h})|^2 = \frac{|F_{\rm obs}(\mathbf{h})|^2}{\Sigma_c(\mathbf{h})} \quad \text{for } \mathbf{h} \text{ centric}, \tag{6}$$

$$|E(\mathbf{h})|^2 = \frac{|F_{\rm obs}(\mathbf{h})|^2}{2\Sigma_a(\mathbf{h})} \quad \text{for } \mathbf{h} \text{ acentric}, \tag{7}$$

so that the expectation value $\langle|E(\mathbf{h})|^2\rangle = 1$ only in the cases for which all assumptions of Wilson statistics are valid.

2.2. Debye's equation and the radial pair distribution

The scattering intensity $I(\mathbf{h})$ from an object consisting of N atoms is given by

$$I(\mathbf{h}) = |F(\mathbf{h})|^2 = F(\mathbf{h})F(\mathbf{h})^* = \textstyle\sum_{i,j} f_i(d^*)f_j(d^*)\exp[2\pi i\mathbf{h}^T(\mathbf{r}_i-\mathbf{r}_j)]. \tag{8}$$

Averaging over the sphere for a given radius d^* gives

$$I(d^*) = \langle I(\mathbf{h})\rangle_{\theta_{d^*}\varphi_{d^*}} = \Big\langle \sum_i f_i^2(d^*)\Big\rangle + \Big\langle \sum_i \sum_{j\neq i} f_i(d^*)f_j(d^*)\cos(2\pi\mathbf{h}^T\mathbf{r}_{ij})\Big\rangle. \tag{9}$$

The average over the spherical angles gives the well known result $\langle\cos(2\pi d^*r_{ij})\rangle = \sin(2\pi d^*r_{ij})/2\pi d^*r_{ij} = \mathrm{sinc}(2\pi d^*r_{ij})$ (the sinc function) and Debye's formula (Debye, 1915) follows,

$$I(d^*) = \sum_i f_i^2(d^*) + \sum_i \sum_{j\neq i} f_i(d^*)f_j(d^*)\,\mathrm{sinc}(2\pi d^*r_{ij}). \tag{10}$$

The scattering intensity is well known to be the Fourier transform of the Patterson function,

$$I(\mathbf{h}) = \mathcal{FT}\{P(\mathbf{r})\} = \int_V P(\mathbf{r})\exp(2\pi i\mathbf{h}^T\mathbf{r})\,\mathrm{d}\mathbf{r}. \tag{11}$$

Averaging over the spherical angle gives the radial intensity

$$I(d^*) = \langle I(\mathbf{h})\rangle_{S_2} = \frac{1}{4\pi}\int_{S_2} I(\mathbf{h})\sin\theta_{d^*}\,\mathrm{d}\theta_{d^*}\,\mathrm{d}\varphi_{d^*} \tag{12}$$

and, with the introduction of the radial Patterson function

$$P(r) = \langle P(\mathbf{r})\rangle_{S_2} = \frac{1}{4\pi}\int_{S_2} P(\mathbf{r})\sin\theta_r\,\mathrm{d}\theta_r\,\mathrm{d}\varphi_r, \tag{13}$$

the intensity can be written as a spherically symmetric Fourier transform of the radial Patterson function,

$$I(d^*) = \int_0^R P(r)\frac{\sin(2\pi d^* r)}{2\pi d^* r} r^2 \, \mathrm{d}r. \qquad (14)$$

The upper bound R simply states that the object (the macromolecule) is finite and that there is no interatomic distance greater than R, $P(r) = 0$ for $r > R$.

For truly equal-atom structures consisting of spherically symmetrical atomic densities, we have $f_i(d^*) = f_j(d^*)$ for all i, j and Debye's equation may be cleanly separated into a purely geometrical term (*interference function* or *structure factor* in small-angle scattering terminology), $S(d^*)$, and scattering term (*form factor*), Form(d^*),

$$I(d^*) = \underbrace{Nf(d^*)^2}_{\mathrm{Form}(d^*)} \underbrace{\left[1 + \frac{1}{N}\sum_i \sum_{j\neq i} \mathrm{sinc}\,(2\pi d^* r_{ij})\right]}_{S(d^*)}. \qquad (15)$$

Division by the scattering term $Nf(d^*)^2$ puts the intensity on a scale independent of the number of atoms and of atom type and so reduces the picture to that of a distribution of point atoms of unit scattering strength. As the sinc function averages out to zero over uniform random position vectors, the expectation for geometric intensity $I(d^*)/Nf(d^*)^2$ in such a case is unity. In effect, the introduction of normalized structure factors attempts to achieve this same factorization into a scattering and a geometric term as best as possible by the replacement of $Nf(d^*)^2$ by $\langle \sum f(d^*)^2 \rangle$ (with correction factors for multiplicity and temperature factors), $|E|^2(d^*) \simeq S(d^*)$, an approximation that breaks down as soon as the individual atomic scattering factors differ too much from each other (*e.g.* heavy atoms), thus causing an overweighting of the corresponding sinc-function contributions.

The summation over all contributing particles in the sinc summation is commonly replaced by a radial distribution function or expressed in the form of a correlation function,

$$S(d^*) = 1 + \frac{1}{N}\sum_i \sum_{j\neq i} \mathrm{sinc}\,(2\pi d^* r_{ij}) = 1 + \frac{1}{N}\int p(r)\mathrm{sinc}\,(2\pi d^* r)\, \mathrm{d}r, \qquad (16)$$

where $p(r)\mathrm{d}r$ gives the number of atoms with a separation within $(r, r + \mathrm{d}r)$. $p(r)$ is known as the distance distribution function or radial pair distribution function (the spherically averaged origin-removed Patterson function). As can be seen, $S(d^*) - 1$ and $p(r)/N$ are related by Fourier transformation. The integral of $p(r)$ over all distances is equal to $N(N-1)$, the number of interatomic distances in the structure. $p(r)/N$ thus gives the average environment per atom. Noting that $p(r)/[N(N-1)]$ is a probability, we can write $S(d^*)$ as $1 + \langle \mathrm{sinc}\,(2\pi d^* r)\rangle$.

The radial pair distribution, $p(r)/N$, of a few selected structures is depicted in Fig. 1. All atoms were considered equal and weighted only by their occupancy. There is a sharp increase at small distances as the first neighbouring atoms are encountered; the major hump in the radial distance distribution is highly shape- and size-dependent, followed by decline at larger distances until a zero value is reached corresponding to the largest interatomic distance within the protein. Fig. 2 shows the $\langle |E| \rangle^2(d^*)$ curves obtained from transforming the radial distance distributions of Fig. 1. The influence of secondary structure on these curves is small but nevertheless significant, as will be shown later.

3. Theoretical $\langle |E|^2 \rangle(d^*)$ profiles and their interpretation

Theoretical profiles for squared normalized structure-factor amplitudes have been computed (taking the PDB isotropic displacement factors and the occupancy into account) for 700 good-quality, low-sequence-similarity, high-resolution protein chains from the Protein Data Bank (Bernstein *et al.*, 1977; Berman *et al.*, 2000). This data set was selected according to the principles described in Hooft *et al.* (1996) with a resolution limit of 2.0 Å. Waters, metals, ligands and riding H atoms were

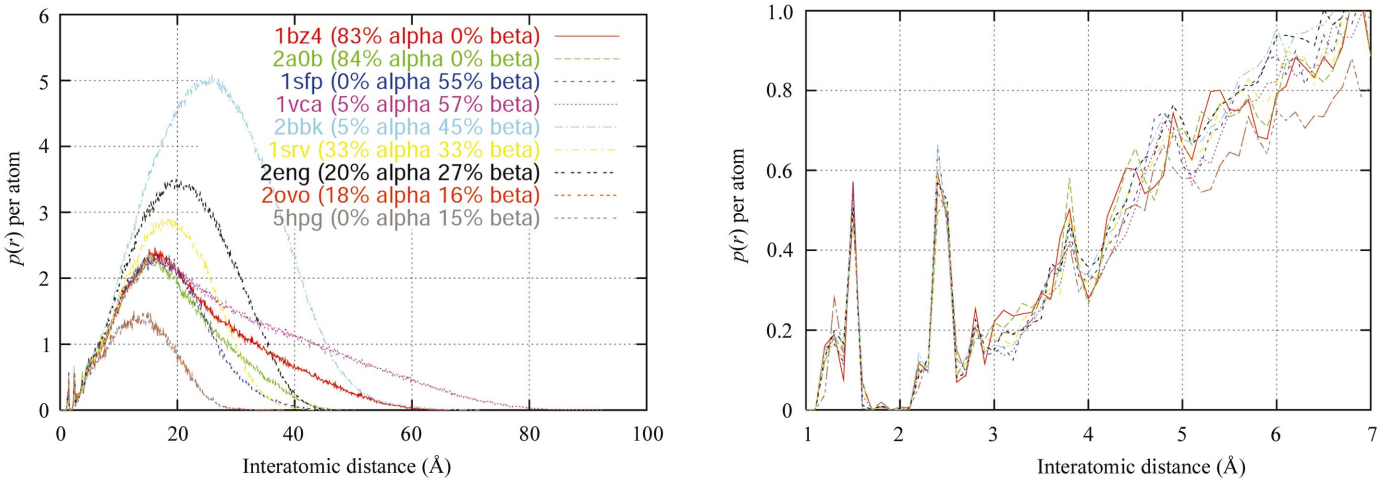

Figure 1
The radial pair distribution function for a few selected protein structures.

excluded from these calculations as only the protein texture was initially of interest. Similar curves have also been computed for a variety of small molecules, proteins and their ligands, DNA and RNA structures. The small-molecule curves will not be discussed here as they add nothing to previous publications in the field of small-molecule crystallography (such as Hall & Subramanian, 1982*a*,*b*). Some examples of DNA with and without protein are presented below (§3.6) after taking a closer look at the protein profiles. For a recent analysis using directly the radial intensity distributions from collected data sets, see Popov & Bourenkov (2003).

3.1. Protein profiles

The most prominent feature of the $\langle|E|^2\rangle(d^*)$ curves is perhaps their striking similarity between very different protein structures, as shown in Figs. 2 and 3 [for reasons of presentation, we depict the natural logarithm of the $\langle|E|^2\rangle(d^*)$ *versus* the resolution in Å in all such plots]. In Fig. 2, a few models covering a wide spread in secondary-structure characteristics are depicted. The similarity between the $\langle|E|^2\rangle(d^*)$ curves is evident in spite of the wide variation in α-helix/β-sheet compositions (assigned using the *DSSP* algorithm; Kabsch & Sander, 1983) and of the globally quite different radial pair-

Table 1
List of extrema in the protein-only $\langle|E|^2\rangle$ profiles.

The inclusion of waters significantly lowers the local minimum of 0.939 at 6.25 Å to a global minimum of 0.386, reduces the local maximum at 4.55 Å from 1.534 to 1.185 and broadens this peak to about 3.5 Å. As mentioned in the main text, these values are dependent on how many waters are used and their assigned *B* factors.

Resolution d (Å)	$\langle|E|^2\rangle$	Extrema
1.12	1.338	Local maximum
1.45	0.808	Local minimum
1.49	0.814	Local maximum
1.69	0.768	Global minimum
2.13	0.947	Local maximum
2.63	0.817	Local minimum
4.55	1.534	Local maximum
6.25	0.939	Local/global minimum
∞	N	Highest point

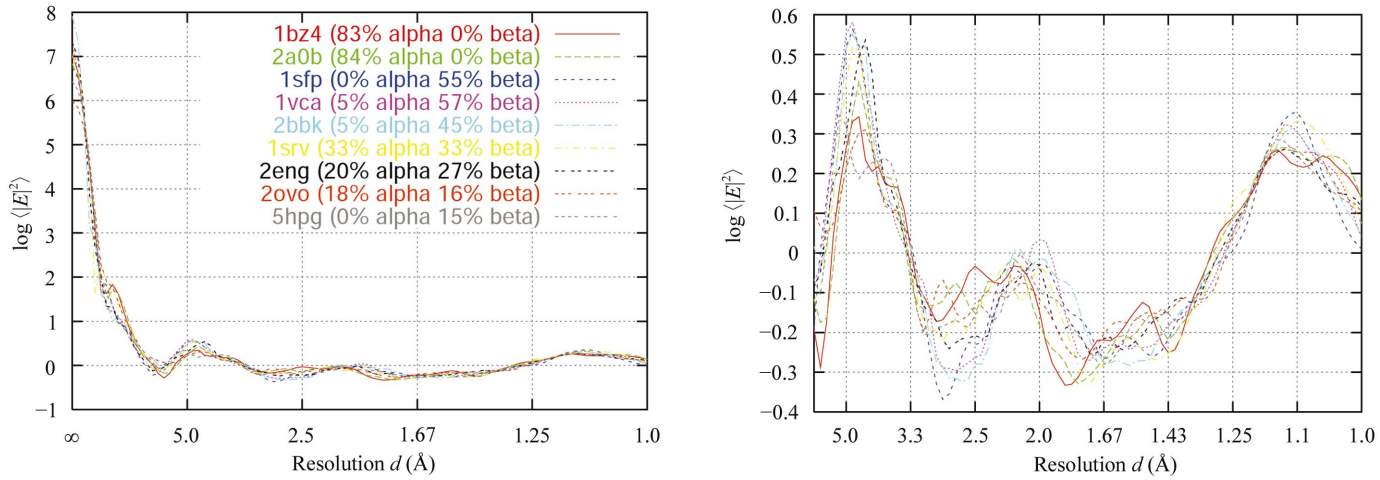

Figure 2
The natural logarithm of the squared normalized structure factors for a few selected protein structures.

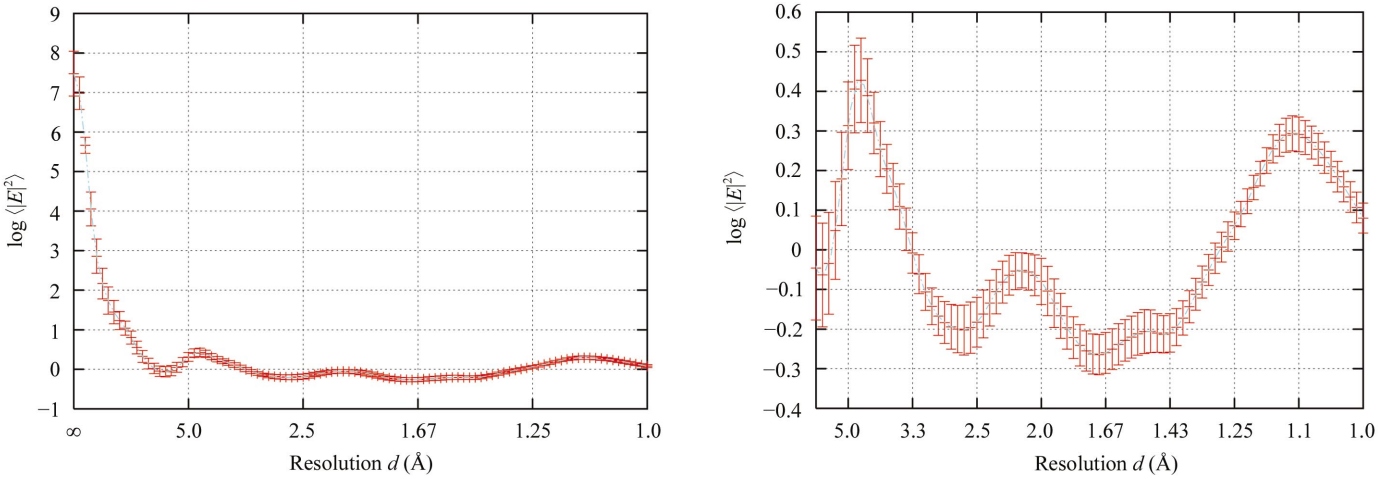

Figure 3
The natural logarithm of the averaged squared normalized structure factors over 700 protein structures with estimated standard deviations. See Table 1 for a list of extrema with corresponding $\langle|E|^2\rangle$ values.

distribution functions (although the finer details, especially in the range up to 6 Å are very similar; see the right-hand plot in Fig. 1).

All structures show a pronounced local maximum at ∼1.1 Å, a smaller local maximum below the expectation value of 1.0 at ∼2.1 Å, a strong local maximum at ∼4.55 Å and a local minimum at ∼6.25 Å. French & Wilson (1978) mention a characteristic minimum at 6 Å and a maximum at 4.5 Å, as have many others, but give no further interpretations. Blessing *et al.* (1996) report a minimum at 6.2 Å and a maximum at 4.4 Å. Their explanation for the minimum is that the 1,3 C^{α} repeats place many atoms near planes with ∼6 Å spacing and destructive interference between these planes is caused by the way they interleave. The maximum at 4.2 Å has been explained by the 1,2 C^{α} repeats placing side-chain atoms on planes approximately separated by 4 Å. This captures a certain amount of truth but requires some further clarifications. (i) An $\langle|E|^2\rangle(d^*)$ profile is the transform of a pair-distribution function and therefore distances rather than planes are the important corresponding features in (one-dimensional) real space to which the (one-dimensional) reciprocal-space features should be matched. (ii) The resolution values, d, at which the $\langle|E|^2\rangle(d^*)$ profiles show features are not the values at which distances (and certainly not planes) should be sought in real space, but rather distances that are sinc-transform-related to these d values. The C^{α} repeats are indeed important, but explaining features in $\langle|E|^2\rangle(d^*)$ profiles at certain d^* values by planes separated by $1/d^*$ is fundamentally incorrect. Planes do not exist in the rotationally averaged one-dimensional space of pair-distribution functions or in d^* space. We give detailed explanations for these $\langle|E|^2\rangle(d^*)$ extrema in terms of atomic distances in the following sections.

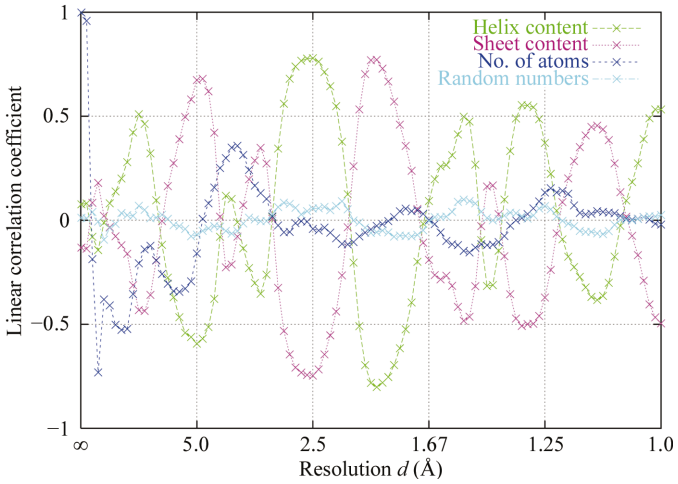

Figure 4
Linear correlation coefficient for α and β content and the size of the protein.

3.2. Correlations with secondary-structure content

Studying the variation of the squared normalized structure-factor amplitudes shell by shell across a large number of structures in conjunction with structural features shows some interesting correlations. Protein secondary-structure assignment has been performed with the *DSSP* algorithm (Kabsch & Sander, 1983) for all 700 structures. For each d^* value, the dependence of the $\langle|E|^2\rangle(d^*)$ values on helix and strand percentage, as well as on the total number of atoms in the structure, was studied by computing a linear correlation coefficient. Fig. 4 shows these dependencies overlaid with the averaged $\langle|E|^2\rangle(d^*)$ over all structures. For a visual estimation of significance, we also depict the correlation coefficient evaluated in the same way for the output of a random-number generator. It is worth noting how far and to what magnitude these correlations extend to high resolution. The correlations in the medium-resolution region are, however, still sufficiently strong for classification purposes, as will be shown below. The helix and strand curves are almost perfect mirror images of each other about the zero-correlation line, thus cancelling out any significant correlation with the total (helix + strand) secondary-structure content.

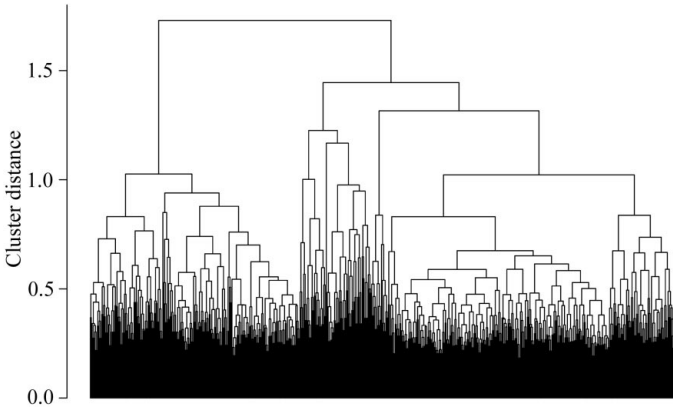

Figure 5
A cluster dendrogram. The clustering of a set of 600 protein structures. There are three main clusters (at cluster distance of about 1.3). The left cluster contains mainly structures with a high strand content (β); the tightly packed centre cluster has mainly structures with a high helix content (α) and the spread right cluster contains mainly α+β structures.

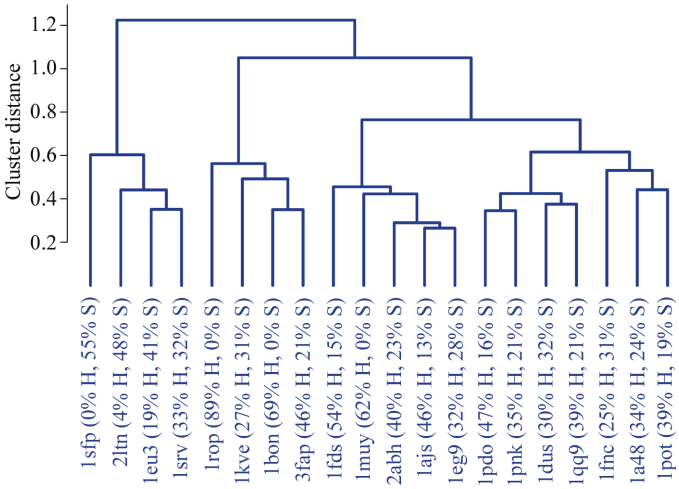

Figure 6
An example of clustering for 20 randomly chosen structures taken from the validation set.

3.3. Clustering

A hierarchical cluster analysis (Murtagh, 1985; Gordon, 1981; Everett, 1974; Ihaka & Gentleman, 1996) was performed on a subset of 600 computed protein $\langle|E|^2\rangle(d^*)$ profiles (Fig. 5). [100 of the 700 set of proteins were kept aside for validation purposes. The structures were used to reproduce a similar clustering from various subsets of data (Fig. 6) and, more importantly, to use the clustering derived with the 600 structures to test the classification power for the 100 structures previously not used directly in the initial clustering process.] Each curve is initially assigned to its own individual cluster. The algorithm then joins the two most similar curves (defined by a simple Euclidean distance metric in this case) and proceeds iteratively until only one cluster remains. This method has the advantage of being solely data-driven in the sense of not imposing a given number of clusters to be found. Plotting the distances between clusters that are to be joined *versus* the individual points/clusters, a *cluster dendrogram* may be constructed (Fig. 7). As hierarchical clustering leads to one final cluster containing all points, visual inspection of the dendrogram is useful in deciding how many relevant clusters should be accepted based on clustering density and the distance to the neighbouring clusters.

To avoid clustering the structures mainly by their number of atoms, only data in the d range 7.0–1.0 Å were used during clustering (see Fig. 7 for a plausible justification of this cutoff). An example of 20 randomly selected structures clustered based on their $\langle|E|^2\rangle(d^*)$ profiles is shown in Fig. 6. Although there are exceptions, a trend can be seen towards clustering of α structures with α structures and of β with β; indeed, three dominant blocks arise in the these clusters that correspond well on average to α, β and $\alpha+\beta$. Using only data corresponding to a resolution of 3.5–7.0 Å, a region commonly said to be the secondary-structure region of the Wilson plot, performs this clustering into secondary structures classes less satisfactorily; indeed, using the 3.5–1.0 Å d range performs a slightly better such classification (although this is of less practical relevance). This performance is judged by building an average helix and strand content per cluster and computing its variance and by counting the number of wrong structures per cluster (data not shown). Such a classification is far from perfect, but an overall trend may be observed. Adjusting the resolution limits can enhance the power and reduce the error rate of classification. A detailed cluster and classification analysis will be presented elsewhere (manuscript in preparation). For a similar classification using wide-angle solution scattering data, see Hirai *et al.* (2002). We have thus shown that although the differences between the three major structure classes (α, β and $\alpha+\beta$) are indeed small, they are sufficient to classify proteins from their computed $\langle|E|^2\rangle(d^*)$ profiles.

3.4. Principal components analysis

After studying the shell by shell correlations, the correlations across shells will now be presented. A principal components analysis (PCA; Fukunaga, 1990; Duda & Hart, 1973) was performed on the $\langle|E|^2\rangle(d^*)$ profiles in the d^* range 0.15–1.0 Å^{-1}. PCA decorrelates features by diagonalizing the covariance matrix of a given set of features. This gives rise to a new set of features which are ordered by decreasing variance, *i.e.* the first few new features account for a large portion of the total variance and are therefore the most useful for classification purposes (see any textbook on pattern recognition for further details; *e.g.* Fukunaga, 1990; Duda & Hart, 1973). The transformation of the $\langle|E|^2\rangle(d^*)$ profiles by the PCA eigenvectors creates as the first feature of the new space a quantity that shows a linear correlation of 0.81 with the helix content and −0.85 with the strand content, thus bringing out in a striking manner the secondary-structure influence on these profiles. The second new feature basically takes the size of the protein into account (Fig. 7). The principal components analysis was however performed with the prior knowledge of the known correlations discussed in the previous section. Performing on the full range from ∞ to 1 Å^{-1} determines different features, of which the first shares 0.997 linear correlation with the size of the protein and only features 8 and 9 show any appreciable correlation with the secondary structure (0.48 helix, −0.36 strand and 0.48 helix, −0.45 strand, respectively). Using data from the 'secondary-structure range' corresponding to 7–3.5 Å, the linear correlation of the first PCA vector is 0.66 to helix and −0.76 to strand content. This means that a simple linear combination of binned $\langle|E|^2\rangle(d^*)$ values in the medium-resolution regime should provide a single number that characterizes the full profile and correlates well the secondary-structure content, thus enabling the same classification as above (α, β, $\alpha+\beta$). In the range corresponding to 3.5–1.0 Å, the linear correlation coefficient is 0.78 and −0.78, respectively, with helix and strand content, as in the case of clustering, indicating that much secondary-structure information resides in this high-resolution range.

3.5. Resolution regimes

The applicability of various theoretical approaches to structure solution and refinement is, as for any optimization

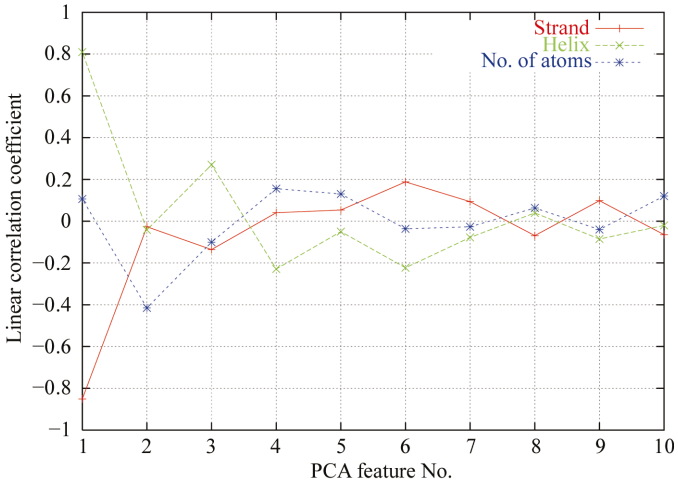

Figure 7
Linear correlation coefficient for α and β content and the size of the protein for the top ten new features from the principal components analysis (linear combinations of the resolutions bins).

procedure, largely a question of the observations-to-parameter ratio. As the desired molecular model from an X-ray diffraction experiment is an atomic one, this imposes the use of additional restraints and constraints to allow the optimization to proceed smoothly with a decreasing observation-to-parameter ratio. One would expect this addition of restraints/constraints to increase smoothly in number with the decrease of resolution. Macromolecular crystallographers have been aware of the existence of a number of resolution boundaries since the earliest days of protein crystallography. We will revisit some of these (software-related) resolution regimes in the light of the various features present in the $\langle|E|^2\rangle(d^*)$ profiles.

3.5.1. Sheldrick's 1.2 Å rule for direct methods. One of the most striking features of the calculated $\langle|E|^2\rangle(d^*)$ profiles is the large peak at approximately 1.1 Å. The emergence of this peak can be understood by recalling that $\langle|E|^2\rangle(d^*)$ and the radial pair-distribution function are related by a spherical Fourier transformation. This peak in $\langle|E|^2\rangle(d^*)$ corresponds to the sinc-function maxima for typical organic bonding distances enhanced by the interference of repetitive features of 1.1 Å separation (starting from this 1.5 Å distance peak) in the radial pair distribution function. Every interatomic distance within a molecule has a corresponding sinc wave in d^* space. These sinc waves show both constructive and destructive interference. The architecture of proteins just happens to produce peaks in the radial pair distribution function that give rise to constructive interference in the region of $d = 1.1$ Å (see also Zwart & Lamzin, 2003). In Morris & Bricogne (2003) we show that this peak is intimately related to the application limit of traditional direct methods and its connection to an approach (Bricogne, 1994, 1995, 1997*a*,*b*) to overcome the limitations thereof.

3.5.2. The empirical applicability rule of 2.3 Å for *ARP*/*wARP*. The next maximum in the $\langle|E|^2\rangle(d^*)$ is at about 2.2 Å. Distances of about 2.5 Å (for instance $C^{\alpha}-C^{\gamma}$, $C^{\beta}-C^{\delta}$, $C_i^{\alpha}-N_{i+1}$) and about 2.8 Å (opposite atoms in six-membered rings, hydrogen-bonded O—N) give rise to a local maximum in this region. This is roughly the resolution needed to be able to resolve the triangle of two successive bonds. Although this is beyond the limit with which atoms can be placed with high accuracy, it is important to reconstruct correctly chemical units the size of the peptide plane. This maximum in the $\langle|E|^2\rangle(d^*)$ profiles at about 2.2 Å is remarkably close to the quoted limit for the application of the automated model-building routine *warpNtrace* (Perrakis *et al.*, 1999) of the *ARP*/*wARP* software suite (Lamzin *et al.*, 2001). *ARP* (Lamzin & Wilson, 1997) itself requires higher resolution to place its atoms with confidence, but once the dummy atoms are more or less in place, *ARP*/*wARP* uses peptide-plane density to recognize main-chain fragments and the introduction of stereochemical restraints of the interpreted fragments to overcome the resolution deficiency. By reducing the dominance of the density criteria and introducing additional longer range geometrical criteria (Morris *et al.*, 2002), this limit has now been broken and version 6.0 successfully builds at 2.5 Å (Morris *et al.*, 2004).

3.5.3. Location of secondary-structure elements to about 5 Å. The assembly of the protein chain into a three-dimensional structure gives rise to structural features commonly classified as α-helices and β-strands/sheets. These features produce characteristic distances within the protein in the order of 4.5–7 Å (α structures typically show peaks in the pair distribution function at about 4.5, 4.9, 5.4, 6.2, 7.3 Å; β structures at 4.8, 6.1, 6.6, 6.9, 7.6 Å). We have attempted to reproduce the peak in $\langle|E|^2\rangle$ at around 4.6 Å with various arrangements of protein chains containing no secondary-structure elements (main-chain backbone angles φ, ψ that do not belong to the α, β core Ramachandran plot regions) but were not successful (data not shown). In terms of introducing random coordinate changes to the initial structure, it can be shown that the peak around 4.6 Å persists to over 2 Å random error (see also Zwart & Lamzin, 2003), while above this value secondary structure becomes exceedingly hard to correctly assign. Upon introducing secondary-element-type structure (adjusting the main-chain backbone angles φ and ψ to fall into the core α, β regions of the Ramachandran plot), a peak emerges in this 4–5 Å regime of the $\langle|E|^2\rangle(d^*)$ profile. We conclude that the broad peak around 4.6 Å may be interpreted as the required resolution to reproduce the arrangement of secondary-structure elements. The finer details of this peak are shown in Fig. 2. The emergence of this peak, as for all the others, can be understood in terms of the sinc-function behaviour of the characteristic distances and the inference with other distance contributions, as described in Morris & Bricogne (2003) and Zwart & Lamzin (2003). Distances typical for secondary-structure elements give rise to a large peak around 4–5 Å. This $\langle|E|^2\rangle(d^*)$ maximum value fits roughly with the limits of programs such as *ESSENS* (Kleywegt & Jones, 1997) and *FFFEAR* (Cowtan & Main, 1998) that successfully place secondary-structure fragments, especially helices, in electron-density maps calculated with data extending to about 5 Å.

3.5.4. 6–7 Å bad data solvent cutoff. In all protein $\langle|E|^2\rangle(d^*)$ curves a prominent minimum can be observed at around 6.3 Å. Beyond this limit the curves are dominated by the shape and size of the protein, implying that the surrounding solvent of the molecule will also contribute to this resolution, as is readily seen from Babinet's principle. The importance of low-resolution data was often overlooked late into the last millennium and these data were not recorded with sufficient care, resulting in corrupted low-resolution Fourier synthesis contributions. In the absence of an adequate solvent correction it proved advantageous to completely leave out these data and to feed refinement programs with heavily weighted parameter restraints to compensate. The suggested cutoff for poor low-resolution data is often in the range 6–7 Å. It must, however, be pointed out that this 'solvent dip' shown in Fig. 3 was reproduced without any waters in the structures and without any kind of solvent modelling. This dip is the consequence of sinc-function contributions arising from the repeats typical for secondary-structure elements. This minimum is greatly enhanced by accounting for the solvent in the calculations as will be shown later. Indeed, Ih ice has

oxygen distances of 4.511 Å (Goto *et al.*, 1990) that produce sinc minima in exactly the same d^* region as the secondary-structure distances.

3.6. Extension to nucleic acids

These analyses have also been carried out for 200 DNA/RNA-only structures from the Protein Data Bank (Bernstein *et al.*, 1977; Berman *et al.*, 2000) (Fig. 8). They too exhibit a remarkably low spread at every resolution, similar to the protein profiles in Fig. 2. However, a number of differences exist. The bonding distances in DNA/RNA exhibit a slightly different distribution compared with proteins, mainly owing to the phosphate group (P—O distances about 1.57–1.61 Å along the main chain and about 1.46 Å for the terminal O atoms in the phosphate group), giving rise to the shifted (compared with the protein profiles) maximum just below 1.1 Å. Contributions to this high-resolution peak and most notably to the second maximum at around 1.7 Å (as compared with about 2.1 Å for proteins) arise from the distances between the stacked purine and pyrimidine side chains of the double-stranded DNA structure. The side chains are stacked at normal distances of about 3.3–3.4 Å, but a range of strong peaks for distances between 3.3 and about 6.0 Å (especially pronounced at around 3.5, 4.3 and 5.2 Å) can be observed between the atoms belonging to any two stacked side chains. The sinc function for distances of 3.5 Å shows a pronounced maximum at d = 1.7 Å and that for distances of 4.3 Å exhibits a minimum at d = 2.5 Å. Note also that the 'solvent dip' at about 6–7 Å is not present. This is related to the slightly different distribution of bonding distances and the reduced number of pronounced peaks in the pair-distribution function for distances of about 4.5–7 Å as seen for protein secondary-structure elements (experimental data do exhibit a dip in this region that indeed arises from the solvent). Similarly, all such minima and maxima can be understood by examining the pair-distribution function (taking into account also the atomic numbers if heavy atoms are present in the structure) and considering the sinc-function contributions and their interferences.

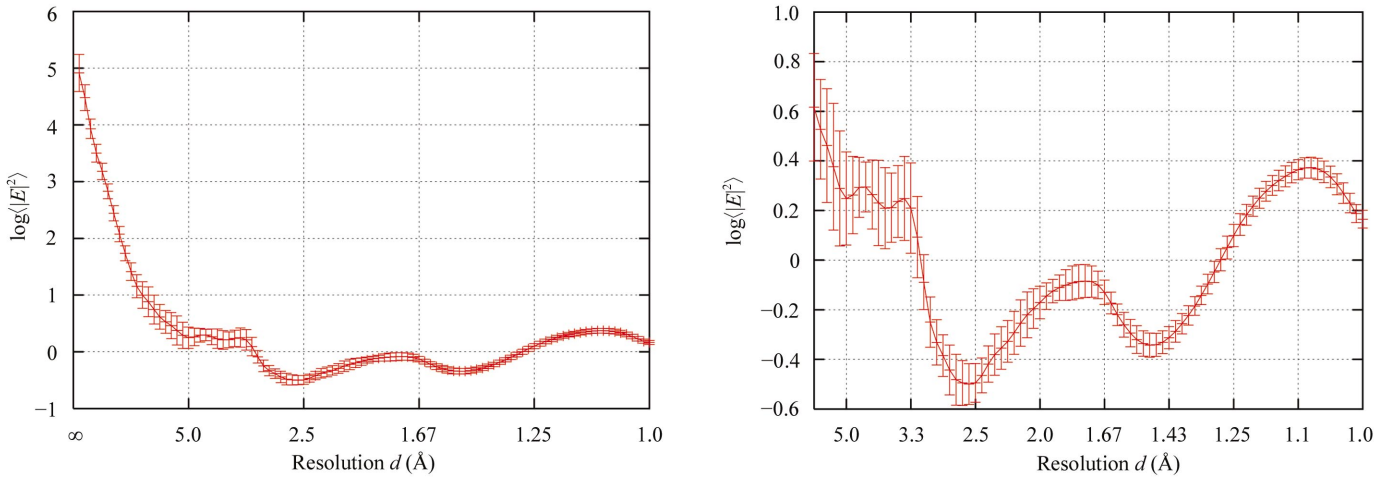

Figure 8
The natural logarithm of the averaged squared normalized structure factors over 200 nucleic acid structures with estimated standard deviations.

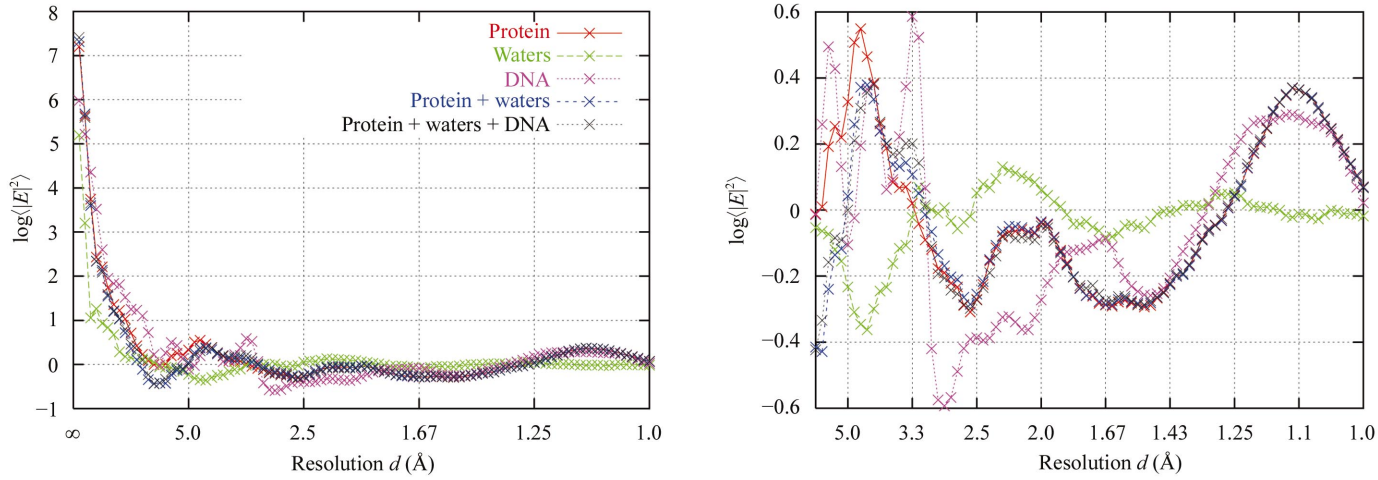

Figure 9
The natural logarithm of the squared normalized structure factors for 1mnn.

We have computed similar curves for nucleic acid and protein structures including ligands. An example with DNA is shown below for which all components of a PDB model (1mnn; Lamoureux *et al.*, 2002) are examined individually and in various combinations (Fig. 9). This model contains 2382 protein atoms, 342 waters and 568 DNA atoms. Unless there is a very significant proportion of DNA or RNA (or heavy atoms) in the model, the protein profile is dominant and dictates the overall trend of the curve. Deviations occur mainly in the 3–5 Å regime as this is the region directly related to the three-dimensional shape of secondary-structure elements that is absent in DNA, RNA and other ligands. As another example, we show PDB model 1a3o which contains 4303 protein atoms, 172 haem-group atoms (including four Fe heavy atoms) as a ligand and 412 waters. Again, we depict all components in various combinations to give an impression for the different contributions (Fig. 10).

All the computations so far are essentially for macromolecules in a vacuum, *i.e.* we have made no attempt to account for the bulk solvent. However, we do discuss this problem in some detail in a later section in connection with scaling.

4. Scaling

4.1. Wilson scaling

Equations (6) and (7) allow measured structure-factor amplitudes to be normalized. Measured data are however not on an absolute scale. It is common to consider an overall scale factor k and an overall temperature factor (a global atomic displacement parameter providing additional resolution-dependent scaling) B for the data. The mapping of the arbitrary observed scale to absolute may therefore be achieved by the multiplication of the measured structure-factor amplitudes by $k\exp(-0.25Bd^{*2})$. As was shown above, an independent uniform random-atom model lacks interatomic interference and therefore the solid-angle averaged squared structure factors should be equal to the summed squared atomic scattering factors. Plotting the logarithm of this ratio (averaged measured squared structure factors over the sum of squared form factors) *versus* d^{*2} should produce values that lie on a straight line of slope $-0.25B$ and shifted by $-\log k$. This or the extension to anisotropic scaling (fit of a tensor β) is the basis of most common attempts to put data on absolute scale. Other methods include origin Patterson peak analysis (Rogers, 1980; Blessing & Langs, 1988) and the K-curve procedure (Karle & Hauptman, 1953). Blessing *et al.* (1996) have shown that if this overall resolution-dependent scale factor is interpreted as an average temperature factor for the structure, then an additional d^*-dependent term arises, as the expectation for the Debye–Waller factor under the under assumption of normally distributed individual B values is $\exp[-0.25(\langle B\rangle - \sigma^2 d^{*2})d^{*2}]$, where $\sigma^2 = \langle(B - \langle B\rangle)^2\rangle$. Although the initial assumptions can be questioned and the behaviour of this result at high resolution is somewhat suspicious, this approach combined with iterative non-linear least squares has been reported to produce reliable normalized structure-factor amplitudes for data extending to about 2.5 Å. These approaches have known problems with data extending little further than about 3 Å owing to the modulation of the straight Wilson line by structural features of macromolecules. A method working to about 5.0 Å was reported in Cowtan & Main (1998) which normalizes a number of scattering curves by the numbers of ordered structures in the unit and attempts to fit these curves to experimental data.

4.2. $\langle|E|^2\rangle$ profile scaling

The now well known and remarkably constant form of the $\langle|E|^2\rangle$ profiles takes into account precisely those variations that can cause problems in Wilson scaling. Their use in scaling would therefore appear to be an attractive way to increase the robustness of current scaling procedures. One problem however has yet to be dealt with.

4.2.1. Water-layer solvent correction. The $\langle|E|^2\rangle(d^*)$ profiles presented so far are all hypothetical in the sense that they were calculated for protein models in a vacuum. This

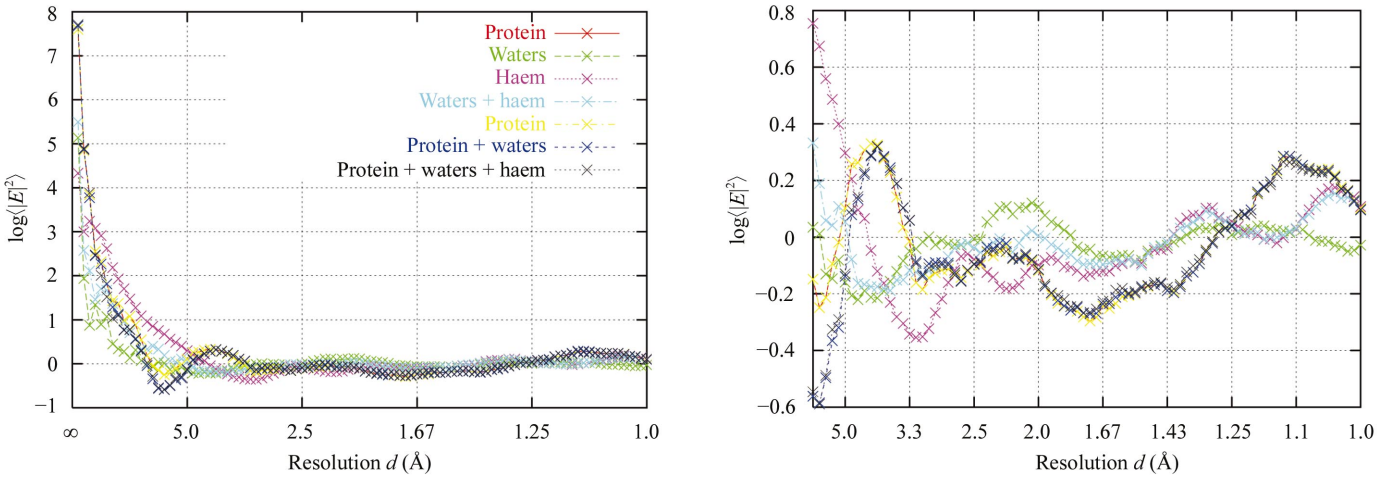

Figure 10
The natural logarithm of the squared normalized structure factors for 1a3o.

allowed a detailed analysis of the correspondence of the actual protein structure and its profile, but should be considered highly suspicious for real data (at least to about 5–6 Å).

To effectively use $\langle|E|^2\rangle(d^*)$ profiles in macromolecular crystallographic software, the contributions of the solvent must be correctly accounted for (Blessing *et al.*, 1996). Methods for crystallographic structure-factor correction and solvent-related techniques may be found in Bricogne (1974, 1976), Phillips (1980), Urzhumtsev & Podjarny (1995), Podjarny & Urzhumtsev (1997), Zhang & Main (1990), Tronrud (1997) and many others. Svergun and co-workers (Svergun *et al.*, 1995; Koch *et al.*, 2003) have developed robust and well tested procedures for modelling the solvent for small-angle scattering applications. We follow a very similar procedure. A solvent layer is placed on an angular grid around the protein surface (this includes cavities) at approximately hydrogen-bonding distance from the protein atoms or bound waters. (Good-quality recent PDB models now commonly have accurate waters, but in general it is advisable to strip especially older deposited structures of their waters as these were often just noise-fitting parameters during refinement.) Babinet's principle is then employed to compute the structure-factor contributions from pseudo-atoms in the structure that have atomic scattering factors corresponding to the displaced solvent. The breadth and depth of the solvent dip in the $\langle|E|^2\rangle(d^*)$ is highly dependent on the choice of parameters such as the number of solvent layers to add around the macromolecule, their shell structure, their atomic displacement factors and their atomic scattering factors (Jiang & Brünger, 1994; Svergun *et al.*, 1998). The radius-dependent shell density model (Schoenborn, 1988; Cheng & Schoenborn, 1990) attempts to account for this behaviour, which was also predicted from molecular-dynamics simulations (Levitt & Sharon, 1988). We also observed a shift towards lower resolution of the minimum with increased solvent modelling and the emergence of a local maximum in the region around 3.5 Å. This observation was also made by Blessing *et al.* (1996). This peak is more pronounced the more solvent there is and the lower the solvent B factor (Blessing *et al.*, 1996). This resolution may therefore be of importance for the estimation of the solvent content and be relevant for phase-improvement techniques. The calculated $\langle|E|^2\rangle(d^*)$ profiles from structures with their modelled water structure in depicted in Fig. 11.

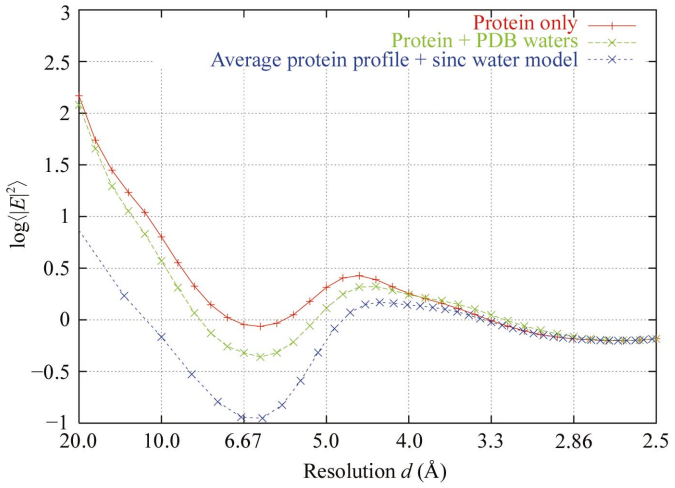

Figure 11
The natural logarithm of the averaged squared normalized structure factors over protein structures including their solvent atoms, using our simple new solvent model with parameters B = 100 Å^2 and $N_{\rm waters} = 3N_{\rm atoms}$ and for 700 protein structures without waters. The latter curve gives a much better prediction of the 'solvent dip' and hence a straighter Wilson plot in Figs. 12 and 13.

It must be stressed that even with adequate solvent modelling [a correct estimate of the extent of the solvent in all directions, a good description either in terms of an atomic structure with appropriate dynamics or a lower parameter description such as in Roversi *et al.* (2000), an appropriate scattering factor description *etc*.] the low-resolution region will be problematic to model within the crystal owing to the proximity of neighbouring molecules that will give rise to modulations of the intensity profile in this region. The better approach may be to calculate $\langle|E|^2\rangle$ from high-quality experimental data measured with great care down to very low resolution (Evans *et al.*, 2000) for a great many different proteins and to average these data, as investigated by Cowtan & Main (1998) and Popov & Bourenkov (2003).

4.2.2. Formulation of the scaling problem. The more structural knowledge that is correctly accounted for in the scaling process, the more the Wilson plot [$\log\langle I\rangle$ *versus* $(d^*)^2$] should resemble a straight line (Main, 1976). This allows the scaling procedure to be formulated as an optimization problem in which the objective function to minimize is the deviation from a straight line. We consider a small set of hypotheses, namely that the structure is dominantly of (i) helical, (ii) β or (iii) α+β content or no clear signal can be detected and we use the average $\langle|E|^2\rangle$ that we obtained from the 700 studied protein structures. We are thus looking for the hypothesis $\mathcal{H}_k$ that maximizes the posterior probability,

$$P(\mathcal{H}_k|\mathcal{D}) \simeq \Lambda_k(\mathcal{D})P(\mathcal{H}_k), \tag{17}$$

in which the likelihood function measures the quality of fit to the best straight line that can be drawn through the Wilson plot under one of the hypotheses mentioned above. In every resolution bin the logarithm of average radial intensity is computed and this value is divided by the logarithm of $\langle|E|^2\rangle(d^*)$ from the profile corresponding to the hypothesis $\mathcal{H}_k$. As the number of individual intensities is typically fairly large (we have >500), the distribution of the bin-average intensity may be assumed normal. We extend this assumption to the distribution of the logarithm of the bin-average intensities as this allows a least-squares fit of the scaling parameters. With

$$\chi^2 = \sum_{i=1}^{N}\left(\frac{y_i - a - bx_i}{\sigma_i}\right)^2 \tag{18}$$

and $y_i = \log\langle F^2_{\rm obs}\rangle_{{\rm shell}\ i}$, $a = \log k$ (where k is the Wilson scale factor), $b = B$ (the Wilson B factor) and $x_i = \sin^2\theta_i/\lambda$, we define the likelihood function as

$$\Lambda_k = \exp\left(-\frac{1}{2}\chi^2\right). \quad (19)$$

The goodness-of-fit can be estimated from the χ^2 distribution with $N-2$ degrees of freedom (Press *et al.*, 1992). The $P(\mathcal{H}_k)$ values can be estimated from the relative frequencies of α, β and $\alpha+\beta$ structures from the PDB or subset thereof.

4.2.3. A new solvent model combined with scaling. If all variable influences on the $\langle|E|^2\rangle$ values could be modelled *via* the introduction of parameters X, the above formalism of trying to obtain the straightest line would give rise to a powerful algorithm for both solvent modelling and scaling. Extending the above equation to include other factors of interest and not only the protein class (which could also be a smooth function of helix and β content given an adequate parameterization of the full profiles), one can write

$$P(X|\mathcal{D}) \simeq \Lambda_X(\mathcal{D})P(X), \quad (20)$$

with the likelihood function as above using the χ^2 distribution.

These parameters X would be adjusted as above such that the resulting curve is as straight as possible. To test this approach, we have developed an extremely simple yet powerful solvent model.

Based on well known oxygen distances from ice-crystal structures (see, for example, Goto *et al.*, 1990), we make a crude approximation by ignoring all protein and water interactions and consider only additional sinc contributions from the most dominant distances (2.760, 4.511 and 7.351 Å for Ih ice), the ratios of which we have fixed (values taken from the intensity distribution; Hura *et al.*, 2003). The number of contributions can be readily extended; indeed, the intensity distribution for water (Hura *et al.*, 2003) could be used directly. Our model has only two parameters that roughly correspond to the number of water atoms to take into account and the isotropic displacement parameter of the water (same for all). We choose these parameters in such a way that the resulting curve is the straightest, *i.e.* by optimizing their values to produce a maximum in the above posterior probability (basically the χ^2 distribution unless a strong prior is used to restrain the values). The algorithm may be described by the following steps.

(i) Read in I and $\sigma(I)$ [or F and $\sigma(F)$] for each hkl.

(ii) Read in the sequence and calculate $\Sigma_2(d^*)$.

(iii) Divide every I by $\varepsilon(\mathbf{h})\Sigma_2(d^*)$ and compute bin averages over the given resolution range.

(iv) Read in the average $\langle|E|^2\rangle$ profile (or many for different secondary structures).

(v) Compute a water-corrected profile consisting of the water-intensity profile (approximated here by three sinc functions) multiplied by a resolution-dependent factor (water B factor) and a scaling parameter (the number of water atoms): two parameters, N_w and B_w.

(vi) Optimize these parameters such that the resulting line in the Wilson plot is the straightest, as judged by the χ^2 likelihood under any given prior knowledge. This fitted straight line gives rise to the Wilson scale and B factor.

Figs. 12 and 13 show two such attempts at modelling the solvent. Despite its simplicity, this approach gave very satisfying results and the optimized parameters even lie in reasonable ranges: we observed values between 90 and 150 Å^2 for the water B factor and the number of water atoms was about three times the number of protein atoms. Considering these values are mainly optimization parameters in a very simple model and were obtained using a non-informative prior, we find they lie in surprisingly reasonable ranges. We are optimistic that an extension of this method may lead to very robust absolute scaling down to about 5 Å and possibly further. Already within the current implementation, we obtain consistent overall Wilson temperature and scale factors down to about 3–3.5 Å (cut data) and acceptable results to about 4.5 Å (cut data). At lower resolution we obtain B factors of the order of 1–5 Å^2 which are still considerably better than the negative values that may otherwise result. The introduction of stronger prior knowledge about the expected B values should increase the robustness further.

4.2.4. Preliminary results. The above approach has been implemented with the crystallographic software development test-bed *BALLS* (Blanc & Morris, 2002).

We show here just two selected examples: PDB codes 1gw9 (Evans & Bricogne, 2002) and 1gwd (Evans & Bricogne, 2002). The data for these structures extend to 1.55 and 1.8 Å, respectively, and are of good quality; therefore, scaling is not a problem. Our intention here is simply to show that the data themselves can correctly select the appropriate $\langle|E|^2\rangle$ profile that is best suited for scaling (see Tables 2 and 3). Note that the resolution range will influence the actual probability values as will the associated errors in each resolution shell: comparing values between data sets should be performed with caution. We use here the vacuum protein profiles but also show first results including our solvent-modelling attempts. The scaling parameters do not vary significantly as enough data exist to provide a robust estimate regardless of proper normalization. The line-fitting quality does however vary as the Wilson curve is straightened by the profile. The actual χ^2 and Λ are highly error-model-dependent and larger error estimates in the resolution bins can produce a seemingly better fit although the scaling parameters may essentially be the same. We have therefore not used the real estimates here, but have quoted the χ^2 and Λ values using in each case the same errors (the set of largest errors from each individual profile). This means that the differences between the quality estimates are smaller than otherwise, but we eliminate the effect of profiles having different σ values owing to the different numbers of structures that were used to calculate them. We will not elaborate further on these issues as the full implementation with anisotropic Wilson B factors and without binned resolution shells is now under way.

The importance of correctly accounting for the effect of the surrounding solvent is vital to extend this approach to lower resolution successfully. In Tables 2 and 3, we have given the parameters using the solvent as described above. It is clearly better than not taking the solvent into account, but further work is needed to better model these influences. The examples

Table 2
Example of automatic profile selection for 1gw9.

This table lists the isotropic scaling parameters for 1gw9 (Evans & Bricogne, 2002) with various profiles used for normalization. The data correctly give preference to the α+β profile from α, β, α+β although the average profile gives clearly better results, probably owing to the better statistics. As a reference the parameters using the profile calculated from the solved structure are given, as is our initial attempt at modelling the solvent effects.

Profile	Wilson B (e.s.d.)	Wilson k (e.s.d)	χ^2_{obs}	$P(\chi^2 > \chi^2_{obs})$
None	12.22 (0.35)	0.260 (0.006)	179.72	1.28×10^{-6}
α	12.64 (0.62)	0.203 (0.009)	136.52	0.0075
β	12.55 (0.62)	0.208 (0.009)	172.54	6.77×10^{-6}
α+β	12.81 (0.62)	0.198 (0.009)	134.33	0.0105
Self	12.45 (0.63)	0.203 (0.009)	101.60	0.4088
Average	12.62 (0.62)	0.202 (0.009)	120.54	0.0696
Average + solvent	13.04 (0.63)	0.190 (0.008)	85.89	0.8230

Table 3
Example of automatic profile selection for 1gwd.

This table lists the isotropic scaling parameters for 1gwd (Evans & Bricogne, 2002) with various profiles used for normalization. The data correctly give preference to the α profile from α, β, α+β, although the average profile gives slightly better results, probably owing to the better statistics. In this case, the errors used in the weighting of χ^2 are probably overestimated.

Profile	Wilson B (e.s.d.)	Wilson k (e.s.d)	χ^2_{obs}	$P(\chi^2 > \chi^2_{obs})$
None	10.7 (1.2)	0.264 (0.016)	47.28	0.790
α	8.4 (1.2)	0.297 (0.019)	39.82	0.950
β	11.5 (1.2)	0.221 (0.014)	48.32	0.757
α+β	10.1 (1.2)	0.252 (0.016)	46.45	0.815
Self	10.07 (1.23)	0.256 (0.016)	33.42	0.993
Average	9.9 (1.2)	0.260 (0.016)	38.01	0.969
Average + solvent	9.8 (1.2)	0.260 (0.017)	34.95	0.988

used here contain triiodide and a large amount of salt that were neglected in our solvent model. If the distances between these atoms were approximately known from concentration measurements, then one could compute an $\langle|E|^2\rangle(d^*)$ profile that could be added to that of the protein in a similar way to that of the solvent.

An alternative approach of averaging experimental Wilson curves as in Cowtan & Main (1998) and Popov & Bourenkov (2003) may be the route to success until better solvent models are developed, although we think our method may have potential and work on further improvement and testing is under way.

Owing to encouraging results (Figs. 12 and 13), the ideas have recently been included in *SHARP* (de La Fortelle & Bricogne, 1997) and *autoSHARP* (Vonrhein & Bricogne, 2004) to provide more robust absolute scaling of lower resolution data sets. In the refinement and model completion modules of *BUSTER* (Bricogne, 1993) the possibility of using non-Wilson prior distributions has existed and been functional since design stage to take into account chemical texture effects in structure-factor statistics. Work is under way and will be presented elsewhere.

5. Discussion

The basic bonding structure of organic molecules and the regular paths (secondary structure) that protein chains follow in three-dimensional space introduce characteristic features in the radial pair-distribution function that give rise to a highly predictable form of the $\langle|E|^2\rangle(d^*)$ profiles. These profiles may be thought of as a (rotationally averaged) reciprocal-space representation of protein texture. In this article, we have investigated the features of macromolecular $\langle|E|^2\rangle(d^*)$ profiles, thereby attempting to link these features to known structural characteristics and also to the limits of application of a few selected software packages. The theoretical $\langle|E|^2\rangle(d^*)$ profiles, despite being the transform of a spherically averaged structure, contain sufficient information to allow a classification into the three major structural groups α, β and α+β to be performed, even at medium resolution. If the secondary

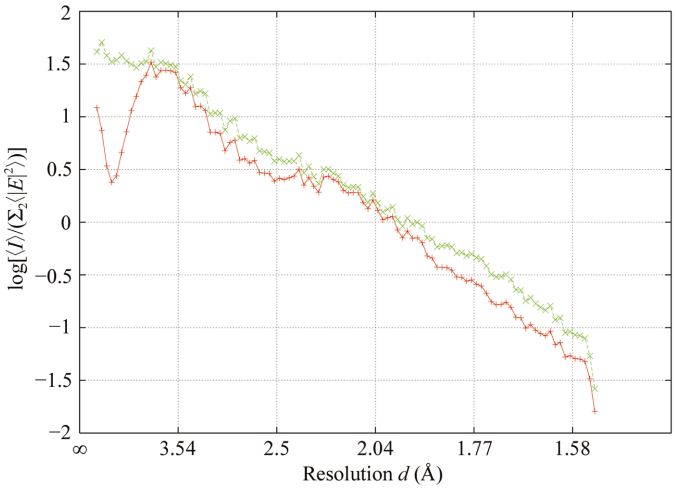

Figure 12
The experimental Wilson plot (with I/Σ_2) for 1gw9 (red) and the corrected Wilson curve (green) by division of the average $\langle|E|^2\rangle$ profile taking solvent effects into account as described in the main text.

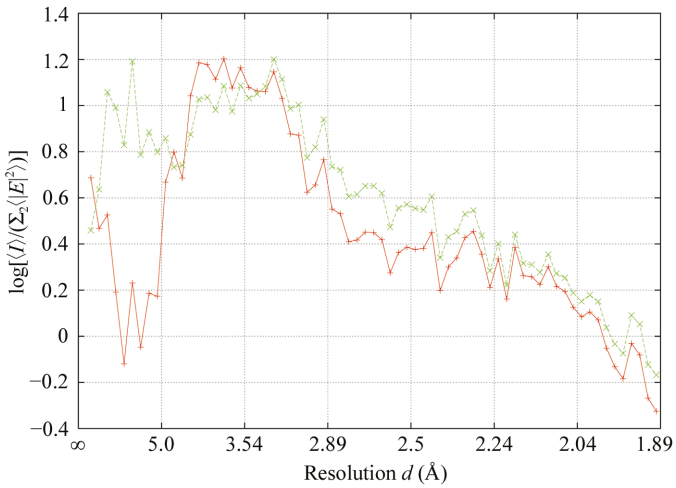

Figure 13
The experimental Wilson plot (with I/Σ_2) for 1gwd (red) and the corrected Wilson curve (green) by division of the average $\langle|E|^2\rangle$ profile taking solvent effects into account as described in the main text.

structure was known reliably beforehand, for instance from homologous structures with high similarity, CD measurements or prediction software, then the data could be scaled using an optimal $\langle|E|^2\rangle(d^*)$ profile for this structure. We have also developed a procedure to allow the profile to be selected automatically. This currently works with three profiles corresponding to the three major protein classes mentioned above, but one could also envision an extended procedure that optimizes the α and β content and chooses the appropriate $\langle|E|^2\rangle(d^*)$ profile to give the straightest line. In many cases, however, experimental error will swamp these subtleties and cause too much blur to enable the fine differences to be used with confidence. Even without secondary-structure-dependent $\langle|E|^2\rangle(d^*)$ profiles, scaling is already made significantly more robust by the use of averaged $\langle|E|^2\rangle(d^*)$ curves and gives more consistent results to lower resolution than without their use. The robustness towards lower resolution will depend mainly on the ability to correctly model the water layer and the bulk-solvent contribution for different molecule shapes and sizes. We have presented here a new solvent-modelling technique that seems to capture well the main features of bulk solvent. This method combined with the averaged squared normalized structure-amplitude profiles has been shown to provide robust absolute scaling in a number of test cases.

We wish to thank Pietro Roversi and Marc Schiltz for detailed discussions and helpful suggestions, and Gwyndaf Evans for high-quality data sets, discussion and advice. RJM thanks Rafael Najmanovich for useful comments on an earlier draft. All authors wish to thank two anonymous referees for valuable critique.

References

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, T. F., Meyer, G. J. B., Brice, E. F. Jr, Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blanc, E. & Morris, R. J. (2002). Unpublished software.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Acta Cryst.* D**52**, 257–266.

Blessing, R. H. & Langs, D. A. (1988). *Acta Cryst.* A**44**, 729–735.

Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.

Bricogne, G. (1976). *Acta Cryst.* A**32**, 832–847.

Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.

Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.

Bricogne, G. (1991). *Acta Cryst.* A**47**, 803–829.

Bricogne, G. (1993). *Acta Cryst.* D**49**, 3–60.

Bricogne, G. (1994). *Likelihood, Bayesian Inference and their Application to the Solution of New Structures*, edited by G. Bricogne & C. Carter, Abstract TRN14, p. 41. American Crystallographic Association.

Bricogne, G. (1995). *ECCC Comput. Chem.* **330**, 449–470.

Bricogne, G. (1997*a*). *Methods Enzymol.* **276**, 361–423.

Bricogne, G. (1997*b*). *Methods Enzymol.* **277**, 14–18.

Cheng, X. & Schoenborn, B. P. (1990). *Acta Cryst.* B**46**, 195–208.

Cowtan, K. & Main, P. (1998). *Acta Cryst.* D**54**, 487–493.

Debye, P. (1915). *Ann. Phys. (Leipzig)*, **46**, 809.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* New York: Wiley–Interscience.

Evans, G. & Bricogne, G. (2002). *Acta Cryst.* D**58**, 976–991.

Evans, G., Roversi, P. & Bricogne, G. (2000). *Acta Cryst.* D**56**, 1304–1311.

Everett, B. (1974). *Cluster Analysis.* London: Heinemann.

French, S. & Wilson, K. (1978). *Acta Cryst.* A**34**, 517–525.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic Press.

Gordon, A. D. (1981).*Classification.* London: Chapman & Hall.

Goto, A., Hondoh, T. & Mae, S. (1990). *J. Chem. Phys.* **93**, 1412–1417.

Hall, S. R. & Subramanian, V. (1982*a*). *Acta Cryst.* A**38**, 590–597.

Hall, S. R. & Subramanian, V. (1982*b*). *Acta Cryst.* A**38**, 598–608.

Harker, D. (1953). *Acta Cryst.* **6**, 731–736.

Hauptman, H. & Karle, J. (1953). *Solution of the Phase Problem I. The Centrosymmetric Crystal*. Dayton, Ohio, USA: American Crystallographic Association.

Hirai, M., Iwase, H., Hayakawa, T., Miura, K. & Inoue, K. (2002). *J. Synchrotron Rad.* **9**, 202–205.

Hooft, R. W. W. & Sander, C. & Vriend, G. (1996). *J. Appl. Cryst.* **29**, 714–716.

Hura, G., Russo, D., Glaeser, R. M., Head-Gordon, T., Krack, M. & Parrinello, M. (2003). *Phys. Chem. Chem. Phys.* **5**, 1981–1991.

Ihaka, R. & Gentleman, R. (1996). *J. Comput. Graph. Stat.* **5**, 299–314.

Jiang. J.-S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Karle, J. & Hauptman, H. (1953). *Acta Cryst.* **6**, 473–476.

Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 208–230.

Koch, M. H. J., Vachette, P. & Svergun, D. I. (2003). *Quart. Rev. Biophys.* **36**, 147–227.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Lamoureux, J. S., Stuart, D., Tsang, R., Wu, C. & Glover, J. N. (2002). *EMBO J.* **21**, 5721–5732.

Lamzin, V. S., Perrakis, A. & Wilson, K. S. (2001). *International Tables for Crystallography* Volume F, edited by M. Rossmann & E. Arnold, pp. 720–722. Dordrecht: Kluwer Academic Publishers.

Lamzin, V. S. & Wilson, K. S. (1997). *Methods Enzymol.* **277**, 269–305.

Levitt, M. & Sharon, R. (1988). *Proc. Natl Acad. Sci. USA*, **85**, 7557–7561.

Luzzati, V. (1955). *Acta Cryst.* **8**, 795–806.

Main, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. Ahmed, K. Huml & B. Sedlaček, pp. 97–105. Copenhagen: Munksgaard.

Morris, R. J. & Bricogne, G. (2003). *Acta Cryst.* D**59**, 615–617.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968–975.

Morris, R. J., Zwart, P., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.

Murtagh, F. (1985). *Multidimensional Clustering Algorithms*. Wuerzburg: Physica-Verlag.

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

Phillips, S. E. V. (1980). *J. Mol. Biol.* **142**, 531–554.

Podjarny, A. D. & Urzhumtsev, A. G. (1997). *Methods Enzymol.* **276**, 641–658.

Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst.* D**59**, 1145–1153.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery B. P. (1992). *Numerical Recipes in C.* Cambridge University Press.

Rogers, D. (1980). *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. Ladd & P. A. Palmer, pp. 82–92. New York: Plenum Press.

Roversi, P., Blanc, E., Vonrhein, C., Evans, G. & Bricogne, G. (2000). *Acta Cryst.* D**56**, 1316–1323.

Schoenborn, B. P. (1988). *J. Mol. Biol.* **201**, 741–749.

Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Stewart, J. M. & Karle, J. (1976). *Acta Cryst.* A**32**, 1005–1007.
Stewart, J. M., Karle, J., Iwasaki, H. & Ito, T. (1977). *Acta Cryst.* A**33**, 519.
Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
Svergun, D., Richard, S., Koch, M. H. J., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 2267–2272.
Tronrud, D. (1997). *Methods Enzymol.* **277**, 306–319.
Urzhumtsev, A. G. & Podjarny, A. D. (1995). *CCP4 Newsl.* **32**, 12–16.
Vonrhein, C. & Bricogne, G. (2004). In preparation.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Wilson, A. J. C. (1950). *Acta Cryst.* **3**, 258–261.
Zhang, K. Y. J. & Main, P. (1990). *Acta Cryst.* A**46**, 377–381.
Zwart, P. H. & Lamzin, V. S. (2003). *Acta Cryst.* D**59**, 2104–2113.