Acta Cryst. (1998). D**54**, 1230–1244

The Two-Dimensional Histogram as a Constraint for Protein Phase Improvement

Alexandra Goldstein† and Kam Y. J. Zhang*

Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA. E-mail: kzhang@fhcrc.org

(*Received 7 October 1997; accepted 30 January 1998*)

Abstract

The joint distribution of electron density and its gradient in a protein electron-density map was examined. This joint distribution was represented by a two-dimensional histogram (2D histogram) of electron-density values and the modulus of the gradient. 16 structures representing distinct protein-fold families were selected to study the dependence of the 2D histogram on resolution, overall temperature factor, structural conformation and phase error. The similarity between the histograms for a pair of structures was measured by correlation coefficient, and the residual provided a measure of the difference. The 2D histogram was found to vary with resolution and overall temperature factor, but was found to be insensitive to structure conformation. The average correlation coefficient between pairs of 2D histograms at three different resolutions examined was 0.90 with a standard deviation of 0.04. The average residual for the same condition was 0.13 with a standard deviation of 0.03. The 2D histogram was also found to be sensitive to phase error. The average correlation coefficient and residual between 2D histograms with 10° phase difference are 0.71 and 0.18, respectively. The variation of the 2D histogram resulting from structure-conformation changes was estimated to be equivalent to that of a 4° phase error. This establishes the minimal phase error that a 2D histogram-matching method could achieve. The conservation of the 2D histogram with respect to structure conformation enables the prediction of the ideal 2D histogram for unknown structures. The sensitivity of the 2D histogram to phase error suggests that it could be used as a target for the density-modification method and also could be used as a figure of merit for phase selection in *ab initio* phasing.

1. Introduction

1.1. *Constraints on the electron density offer a means of phase retrieval*

The crystallographic phase problem is indeterminate given only the structure-factor amplitudes. It is only through knowledge of the chemical or physical properties of the electron density that the phases can be retrieved. Characteristic features of the correct electron density can often be expressed as mathematical constraints on the density function and thereby on the structure-factor phases. In favorable cases, these constraints are sufficient to determine the phases directly, which gives rise to direct methods.

Direct methods for small molecules rely upon the availability of X-ray diffraction data at atomic resolution (Hauptman, 1986; Karle, 1986). It exploits the fact that the electrons scatter the X-rays so that electron density values cannot be negative (positivity). There are discrete atomic peaks at atomic resolution, therefore the density around the peak should assume the shape of an atom (atomicity). However, protein crystals rarely diffract to atomic resolution owing to the special properties of proteins such as the flexibility of the peptide chain and the large solvent content. Consequently, the positivity and atomicity constraints are no longer strictly valid for proteins. Methods that might solve the protein phase problem must be able to deal with non-atomic resolution diffraction. Thus, constraints that are valid at non-atomic resolution are crucial to the success of any methods for protein phasing.

1.2. *Some constraints exploited for protein phase improvement*

For the majority of protein crystals with non-atomic resolution diffraction, the available constraints at our disposal are not sufficient to enable a unique solution to the protein phase problem (Baker, Krukowski *et al.*, 1993). Most current methods are aimed at improving the initial phases and extending them to the full resolution of the diffraction data.

The special features of protein crystals, such as low-resolution diffraction pattern, large unit-cell size, large variation of thermal motion and high solvent content, constitute the major challenges faced in protein phasing. However, these unique features have also been exploited in various methods for phase refinement and extension. For example, solvent flattening exploits the fact that the solvent region of the electron density is flat at medium resolution due to high thermal motion of the atoms and disorder of the solvent (Wang, 1985). Histogram matching utilizes the structural independence of

† Present address: iCat Corporation, 1420 Fifth Avenue, Suite 1800, Seattle, WA 98101, USA.

ISSN 0907-4449

the electron-density distribution to improve the phases by bringing the distribution of electron-density values of a given map to that of an ideal map (Zhang & Main, 1990*a*). Sayre's equation is used to restrain the local shape of the electron density (Sayre, 1952). Molecular averaging forces the electron density at equivalent positions to be equal when there are multiple copies of the same molecule in the asymmetric unit (Bricogne, 1976). Electron-density skeletonization imposes main-chain connectivity in the electron density which is characteristic of protein molecules (Baker, Bystroff *et al.* 1993; Bystroff *et al.*, 1993; Wilson & Agard, 1993).

The combination of all available constraints yields a larger constraint space and reduces the ambiguity of phase values, since each constraint represents different characteristic features in the electron density. The constraints from Sayre's equation, solvent flattening, histogram matching, molecular averaging and map connectivity were combined for phase refinement and extension in the *SQUASH* method (Zhang, 1993; Zhang *et al.*, 1997; Zhang & Main, 1990*b*). It was shown that there was synergism between these constraints, and that the simultaneous application of the constraints yielded the most powerful method for phase improvement.

1.3. *More constraints are needed to solve the protein phase problem*

The effectiveness of a phase-improvement method relies on the phasing power of the constraints and the number of independent constraints used (Arnold & Rossmann, 1986). The phasing power of a constraint depends on the number of density points affected and the magnitude of the changes imposed on the electron density. It also depends on the physical nature and accuracy of the constraint and how rigorously the constraint is applied. The refinement and extension could be initiated from a lower resolution as more constraints are employed. A method to solve the macromolecular phase problem *ab initio* would exist when refinement and extension could be initiated from randomly generated phases. The goal of this work is to find new characteristic features of the electron density, and to combine them with existing constraints to create more powerful methods for phase improvement. Specifically, we have examined the joint probability distribution of the electron density and its gradients.

The density histogram of a map is the probability distribution of the electron-density values. The ideal density histogram has provided a constraint on the electron-density distribution which can be used to improve phases by histogram matching, which seeks to bring the distribution of electron-density values of a given map to that of an ideal map (Harrison, 1988; Lunin, 1988; Zhang & Main, 1990*a*).

The electron-density histogram specifies not only the permitted values of the electron density but also their frequencies of occurrence (distribution). This distribution contains structural information about the underlying protein structure, such as the type of atoms and their packing. Proteins consist of mostly C, N, O and a few S atoms, and these atoms are certain characteristic distances apart. The atoms are packed together in protein structures and the packing density is relatively independent of the structure conformation (Matthews, 1968, 1974). The distribution of atomic types and the distances and angles between different atomic types are all very similar among different structures. The difference in structural conformation mainly arises from the dihedral angles of each residue. Since the density histogram discards the positional information and encodes only the distribution of the electron-density values, it is therefore independent of structural conformation. This means the density histogram is predictable and can be used as a constraint for phase improvement.

The density histogram is degenerate, however, in encoding the structural information. Drastically different structures can have the same density distribution. Moreover, the stereochemical features in protein structures are not all captured in the density histogram. Since the density histogram only accounts for the value at a given point and ignores its neighboring environment, any information about the neighborhood of a point will be complementary to the density histogram. One obvious measure of the environment of the neighboring points is the gradient, since it reflects the change of the density value within a local region of a given point. It is based on this reasoning that we have examined the joint distribution of density and gradient, with the aim of finding a more discriminating constraint that could be used for phase improvement.

2. Methods

There are two criteria that a constraint should satisfy in order to be useful for phase improvement.

2.1.1. *Predictability*. The constraint should be structure independent. The less variation the constraint has from structure to structure, the higher potential the constraint will have for phasing.

2.1.2. *Sensitivity to phase errors*. The constraint should vary with phase errors. The more sensitive to phase errors the constraint is, the higher the phasing power of this constraint.

The 2D histogram must be independent of structural conformation if it is to be predictable. It is known that the one-dimensional electron-density histogram is dependent on resolution and the overall B factor (Zhang & Main, 1990*a*). Therefore, we should first examine the dependence of the 2D histogram on resolution and overall B factor. This is to identify and remove factors that affect the 2D histogram. The search for these factors does not have to be exhaustive, provided that the 2D histogram is independent of

structural conformation after they are removed. Those factors have to be obtainable without the knowledge of the structural conformation. We then examine the dependence of the 2D histogram on structural conformation. This is to establish whether the 2D histogram is predictable. We finally examine the dependence of the 2D histogram on phase error to estimate its potential phasing power.

In this section, we will first illustrate how the density gradients are calculated. We will then describe how the joint distribution of density and gradient is obtained. Finally, the methods used to measure the similarity of the density and gradient histograms are described.

2.2. *The calculation of density gradients*

Given the electron density ρ at position (x, y, z) as a Fourier transform of structure factors $F(hkl)$,

$$\rho(xyz) = (1/V)\sum_{hkl} F(hkl)\exp[-2\pi i(hx + ky + lz)], \tag{1}$$

where (x, y, z) are the fractional coordinates along the crystal axes (**a**, **b**, **c**) and (hkl) are the indices of the structure factors, the gradients along each of the three crystal axes are calculated as

$$\begin{aligned}\frac{\partial\rho(xyz)}{\partial x} &= -(2\pi i/V)\sum_{hkl} hF(hkl)\exp[-2\pi i(hx + ky + lz)]\\ &= \nabla x,\\ \frac{\partial\rho(xyz)}{\partial y} &= -(2\pi i/V)\sum_{hkl} kF(hkl)\exp[-2\pi i(hx + ky + lz)]\\ &= \nabla y,\\ \frac{\partial\rho(xyz)}{\partial z} &= -(2\pi i/V)\sum_{hkl} lF(hkl)\exp[-2\pi i(hx + ky + lz)]\\ &= \nabla z.\end{aligned} \tag{2}$$

The modulus of the gradient, g, is then determined from

$$g^2 = |\nabla\rho|^2 = \nabla\cdot\nabla = (\nabla x\nabla y\nabla z) \times \begin{pmatrix} a^2 & ab\cos\gamma & ac\cos\beta \\ ba\cos\gamma & b^2 & bc\cos\alpha \\ ca\cos\beta & cb\cos\alpha & c^2 \end{pmatrix}\begin{pmatrix}\nabla x\\ \nabla y\\ \nabla z\end{pmatrix}, \tag{3}$$

where a, b, c and α, β, γ are the unit-cell parameters.

2.3. *The accumulation of two-dimensional density and gradient histograms*

The 2D histogram of density and gradient is the joint distribution of the electron-density value and its gradient in the protein region of the unit cell. It provides a global description of the electron-density map and all spatial information is discarded. Although the electron density and its gradient can be calculated analytically from structure factors by Fourier transforms, their joint probability distribution cannot be obtained easily by analytical methods. Therefore, a numerical approach is adopted to derive the joint probability distribution of electron density and gradients. The joint distribution of density and gradient, $P(\rho, g)$, is defined as

$$P(\rho, g) = n(\rho \pm \Delta\rho, g \pm \Delta g)/N, \tag{4}$$

where $n(\rho \pm \Delta\rho, g \pm \Delta g)$ is the number of grid points in the protein region of the map that have the density $\rho \pm \Delta\rho$ and gradient $g \pm \Delta g$, and N is the total number of grid points in the protein region of the map. The density and gradient are divided into 200 bins, *i.e.* $\Delta\rho = (\rho_{\max} - \rho_{\min})/200$ and $\Delta g = (g_{\max} - g_{\min})/200$. Note that the projection of the 2D histogram, $P(\rho, g)$, along g or ρ gives the one-dimensional histograms, $P(\rho)$ or $P(g)$ respectively, *i.e.* $P(\rho) = \int P(\rho, g)\,\mathrm{d}g$ and $P(g) = \int P(\rho, g)\,\mathrm{d}\rho$.

The 2D histogram is obtained through the following protocol.

(i) Calculate the structure-factor amplitudes and phases from the atomic coordinates of a selected structure by inverse Fourier transforms.

(ii) Modify each structure factor, F, to F' by the removal of the overall B factor, $F' = F\exp(Bs^2)$, where s is the reciprocal-space vector. The overall temperature (B) factor is calculated from the structure-factor amplitudes and the unit-cell content of the crystal using Wilson statistics (Wilson, 1949). The modified structure factor corresponds to that of a stationary atom.

(iii) Calculate the electron-density map according to one of the following recipes, based on the factors examined: (*a*) using structure factors to different resolutions to examine the effect of resolution on the 2D histogram, (*b*) using structure factors modified by various overall B factors to examine the effect of atomic thermal motion on the 2D histogram, (*c*) using structure factors from different structures at the same resolution and with the same overall B factor to examine the conformation dependence of the 2D histogram or (*d*) using structure factors with random phase errors added to examine the effect of phase error on the 2D histogram.

(iv) Calculate the density gradient along each crystal axis using equation (2) and calculate the modulus of the gradient using equation (3).

(v) Calculate the molecular envelope of the electron-density map by the reciprocal-space convolution method (Leslie, 1987; Wang, 1985).

(vi) Accumulate the occurrences of density and gradient values in the protein region of the map inside the molecular envelope. Calculate the joint distribution by dividing the number of the occurrences by the total number of grid points in the protein region.

Table 1. *Correlation coefficients and residuals between 2D histograms of fibroblast growth factor at resolutions between 1.6 and 4.0 Å*

Each number in the tables represents the correlation between 2D histograms of two resolutions shown in the first row and the first column. Only the correlation coefficients in upper half of the diagonal are shown since the table is symmetric.

(*a*) Correlation coefficients. Mean = 0.735, variance = 0.179.

Resolution (Å)	4.00	3.42	3.00	2.67	2.40	2.18	2.00	1.85	1.71	1.60
4.00	1.000	0.826	0.629	0.530	0.422	0.390	0.380	0.374	0.371	0.320
3.42	—	1.000	0.873	0.767	0.655	0.607	0.598	0.596	0.593	0.543
3.00	—	—	1.000	0.897	0.808	0.748	0.736	0.729	0.724	0.689
2.67	—	—	—	1.000	0.876	0.826	0.810	0.797	0.792	0.762
2.40	—	—	—	—	1.000	0.888	0.880	0.868	0.862	0.840
2.18	—	—	—	—	—	1.000	0.903	0.901	0.896	0.882
2.00	—	—	—	—	—	—	1.000	0.914	0.911	0.899
1.85	—	—	—	—	—	—	—	1.000	0.920	0.913
1.71	—	—	—	—	—	—	—	—	1.000	0.917
1.60	—	—	—	—	—	—	—	—	—	1.000

(*b*) Residual. Mean = 0.257, variance = 0.118.

Resolution (Å)	4.00	3.42	3.00	2.67	2.40	2.18	2.00	1.85	1.71	1.60
4.00	0.000	0.244	0.373	0.433	0.468	0.474	0.473	0.470	0.471	0.484
3.42	—	0.000	0.195	0.288	0.342	0.359	0.361	0.359	0.360	0.374
3.00	—	—	0.000	0.156	0.217	0.249	0.259	0.266	0.271	0.283
2.67	—	—	—	0.000	0.142	0.176	0.194	0.210	0.219	0.228
2.40	—	—	—	—	0.000	0.136	0.150	0.169	0.178	0.185
2.18	—	—	—	—	—	0.000	0.132	0.141	0.150	0.155
2.00	—	—	—	—	—	—	0.000	0.129	0.134	0.139
1.85	—	—	—	—	—	—	—	0.000	0.126	0.128
1.71	—	—	—	—	—	—	—	—	0.000	0.126
1.60	—	—	—	—	—	—	—	—	—	0.000

2.4. *The measurement of similarity between 2D histograms*

In order to examine the dependence of the 2D histogram on resolution, overall B factor, structure conformation and phase error, we used the correlation coefficient and residual to quantify the similarity and difference between a pair of 2D histograms.

2.4.1. *Correlation coefficient*. The similarity between any two 2D histograms is measured by the correlation coefficient, which is defined as

$$C_{kl} = \left\{ \sum_{j=1}^{n}\sum_{i=1}^{m}[P_k(\rho_i, g_j) - \overline{P_k(\rho_i, g_j)}] \times [P_l(\rho_i, g_j) - \overline{P_l(\rho_i, g_j)}] \right\} \div \left(\left\{ \sum_{j=1}^{n}\sum_{i=1}^{m}[P_k(\rho_i, g_j) - \overline{P_k(\rho_i, g_j)}]^2 \times \sum_{j=1}^{n}\sum_{i=1}^{m}[P_l(\rho_i, g_j) - \overline{P_l(\rho_i, g_j)}]^2 \right\}^{1/2} \right). \quad (5)$$

where $P_k(\rho_i, g_j)$ and $P_l(\rho_i, g_j)$ represent the 2D histograms k and l at a given point (ρ_i, g_j). $\overline{P_k(\rho_i, g_j)}$ and $\overline{P_l(\rho_i, g_j)}$ represent the average values in the 2D histograms.

2.4.2. *Residual*. The difference between any two 2D histograms is measured by the residual which is defined as

$$R_{kl} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m}[|P_k(\rho_i, g_j) - P_l(\rho_i, g_j)|]}{\sum_{j=1}^{n}\sum_{i=1}^{m}\left[|P_k(\rho_i, g_j)| + |P_l(\rho_i, g_j)|\right]}, \quad (6)$$

where $P_k(\rho_i, g_j)$ and $P_l(\rho_i, g_j)$ represent the 2D histograms k and l at a given point (ρ_i, g_j). $|P_k(\rho_i, g_j)|$ and $|P_l(\rho_i, g_j)|$ represent the modulus of $P_k(\rho_i, g_j)$ and $P_l(\rho_i, g_j)$, respectively.

3. Results

The 2D histograms from a representative selection of protein structures were systematically examined to reveal the dependence of 2D histograms on resolution, overall temperature (B) factor, structural conformation and phase error.

3.1. *The resolution dependence*

Resolution reflects the amount of overlap between neighboring atoms. The overlap of electron density between atoms will affect both the density and gradient distribution and, therefore, the 2D histogram. To

determine the extent of the dependence of the 2D histograms on the resolution, we compared the 2D histograms of single structures at ten different resolutions ranging from 1.6 to 4.0 Å. This resolution range was selected based upon the typical resolutions of the experimentally derived multiple isomorphous replacement (MIR) phases that require further refinement or extension. The sampling of resolution is in equal steps of reciprocal-space vector length (the inverse of resolution). One example, using fibroblast growth factor (FGF), is presented here. The 2D histograms of FGF at three selected resolutions, 1.6, 2.2 and 4.0 Å, are shown in Fig. 1(*a*). In order to facilitate the visual inspection of the variation of the 2D histogram, it was projected along the gradient or density to produce the one-dimensional density or gradient histograms, which are shown in Figs. 1(*b*) and 1(*c*), respectively. The resolution dependence of the 2D histograms was quantified by calcu-

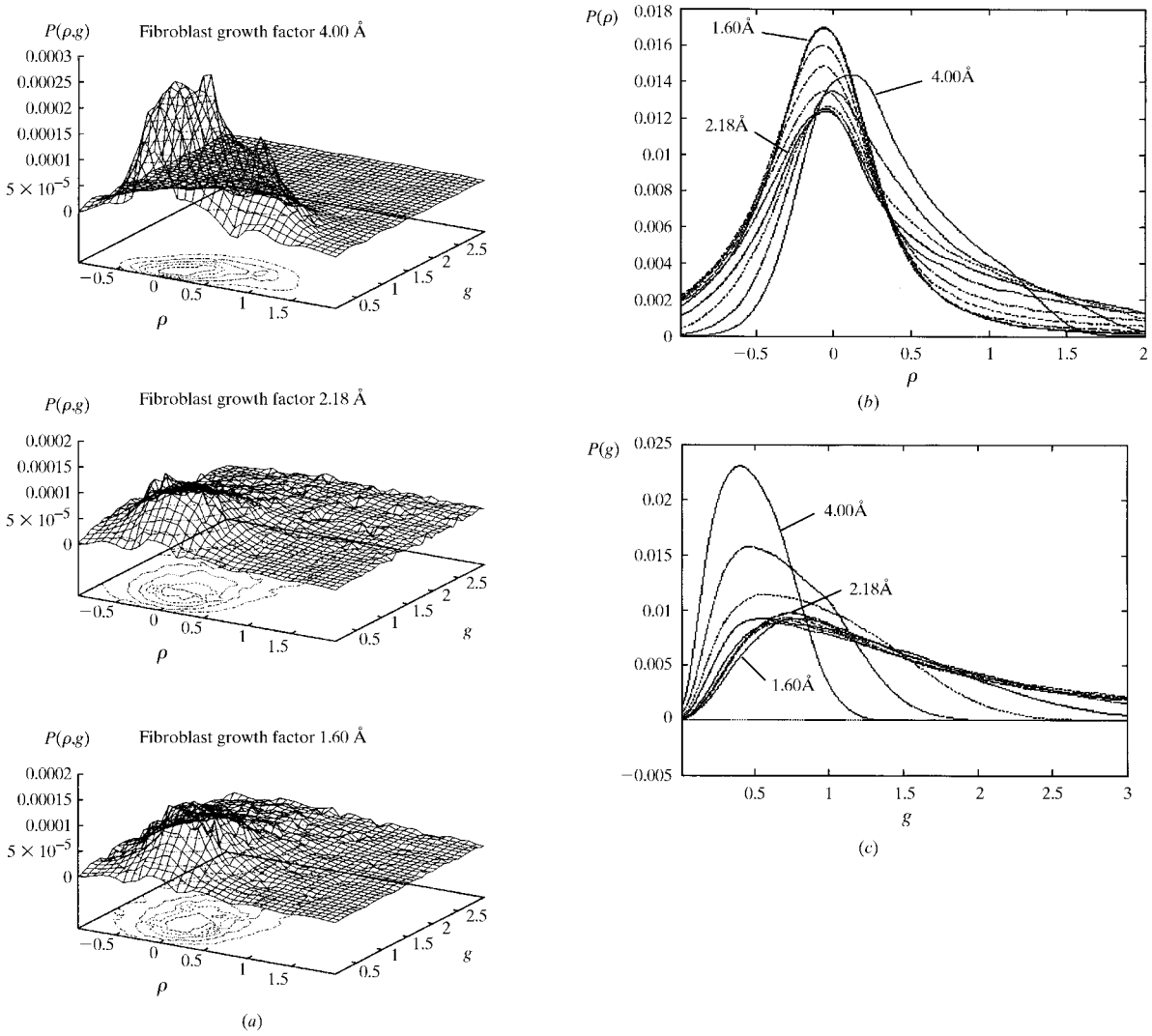

Fig. 1. (*a*) 2D histograms of fibroblast growth factor at 1.6, 2.18 and 4.0 Å resolution. The axes marked by ρ, g and $P(\rho, g)$ represent the electron density, modulus of the gradient and the joint probability distribution of electron density and gradient, respectively. The unit of the electron density is e Å^{-3} and the unit of the gradient is e Å^{-4}. The $P(\rho, g)$ was normalized such that the total area under the surface equals one. All the 2D histograms are plotted on the same scale for comparison. The $P(\rho, g)$ is represented as a mesh surface. A contour of $P(\rho, g)$ is also shown in the figure as dashed lines. The 2D histogram is sensitive to resolution changes. (*b*) One-dimensional density histogram of fibroblast growth factor at resolutions from 1.6 to 4.0 Å. The axes ρ and $P(\rho)$ represent the electron density and its probability respectively. Only three of the ten histograms are labeled for clarity. The rest can be identified by following the trend of the changes with resolution. The map at high resolution has not only more high densities but also more low densities. (*c*) One-dimensional gradient histogram of fibroblast growth factor at resolutions from 1.6 to 4.0 Å. The axes g and $P(g)$ represent the electron-density gradient and its probability, respectively. Only three of the ten histograms are labeled for clarity. Following the trend of the changes with resolution can identify the rest. The map at high resolution has more steep gradients.

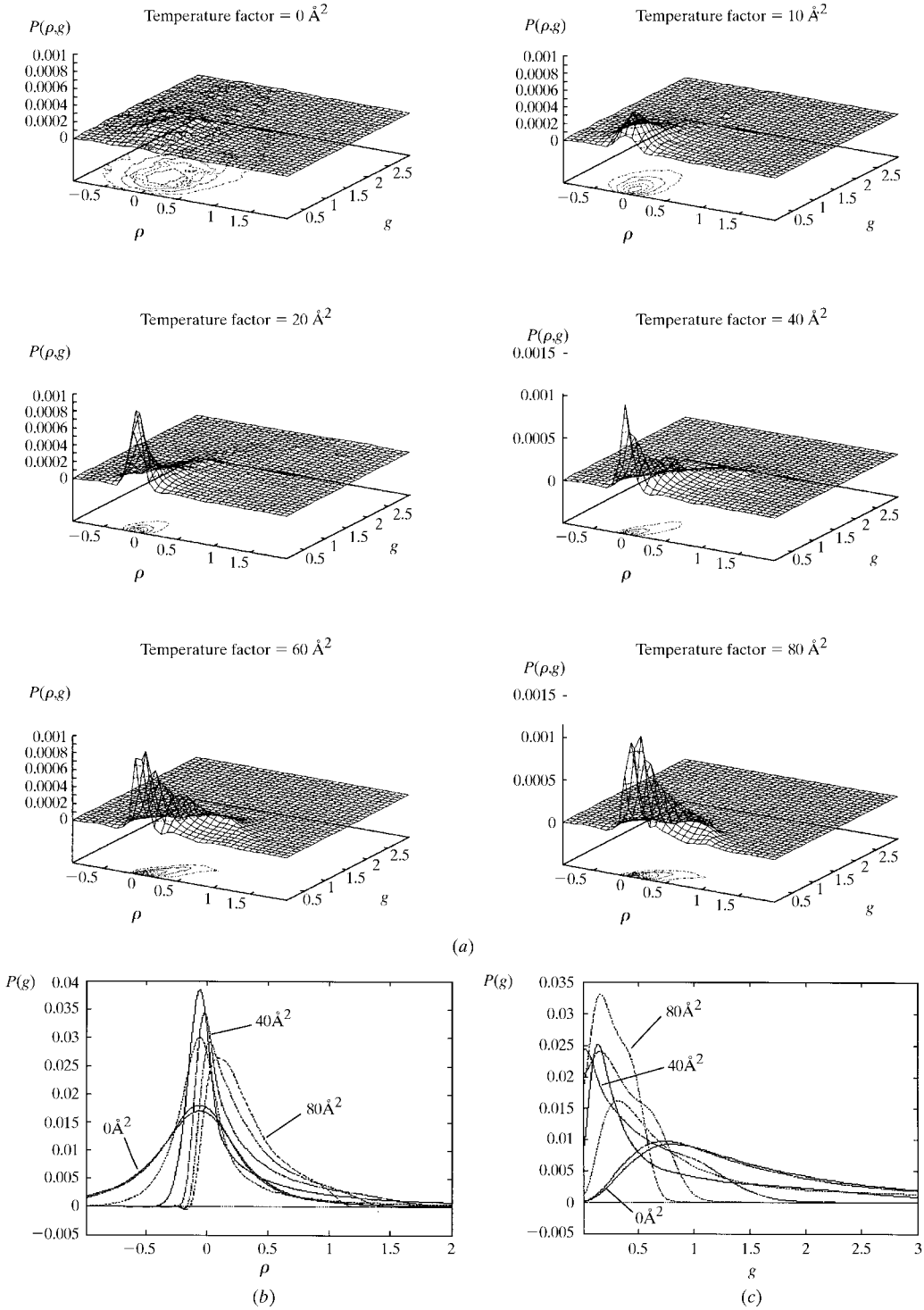

Fig. 2. (*a*) 2D histograms of fibroblast growth factor at 1.6 Å with *B* factors ranging from 0 to 80 Å^2. The 2D histograms from these maps with various *B* factors are drawn on the same scale for comparison. The 2D histogram is clearly dependent on the *B* factor. There are more intermediate densities with low gradients as the *B* factor increases. (*b*) One-dimensional density histogram of fibroblast growth factor at 1.6 Å with *B* factors at 0, 1, 10, 20, 40, 60 and 80 Å^2. Only three of the seven density histograms are labeled for clarity. The effect of *B* factor on density histograms is similar to that of resolution. The maps with large *B* factors have fewer high densities and also fewer low densities. (*c*) One-dimensional gradient histogram of fibroblast growth factor at 1.6 Å with *B* factors at 0, 1, 10, 20, 40, 60 and 80 Å^2. The effect of *B* factor on gradient histograms is also similar to that of resolution. There are fewer and fewer high gradients and more and more low gradients as the *B* factor increases.

Table 2. *Correlation and residual of 2D histograms of FGF at 1.6 Å with B factors ranging from 0 to 80 Å^2*

(*a*) Correlation. Mean = 0.378, variance = 0.288.

Temperature factor (Å^2)	0	1	10	20	40	60	80
0	1.000	0.920	0.568	0.205	0.101	0.043	−0.002
1	—	1.000	0.632	0.246	0.129	0.065	0.013
10	—	—	1.000	0.695	0.413	0.264	0.139
20	—	—	—	1.000	0.713	0.459	0.248
40	—	—	—	—	1.000	0.788	0.484
60	—	—	—	—	—	1.000	0.813
80	—	—	—	—	—	—	1.000

(*b*) Residual. Mean = 0.447, variance = 0.148.

Temperature factor (Å^2)	0	1	10	20	40	60	80
0	0.000	0.125	0.273	0.394	0.541	0.607	0.636
1	—	0.000	0.255	0.384	0.535	0.603	0.633
10	—	—	0.000	0.255	0.457	0.558	0.608
20	—	—	—	0.000	0.348	0.499	0.577
40	—	—	—	—	0.000	0.324	0.484
60	—	—	—	—	—	0.000	0.286
80	—	—	—	—	—	—	0.000

lating the correlation coefficient (Table 1*a*) and residual (Table 1*b*) between 2D histograms of various resolutions using equations (5) and (6), respectively.

The resolution dependence of the 2D histogram can be clearly seen from Fig. 1(*a*). At low resolution, most of the grid points have low density and low gradient and give rise to the high peak in the figure. As the resolution increases, the 2D histogram becomes flatter and there are more grid points in the map that have high-density and high-gradient values. A higher resolution means less overlap of atoms, which gives rise to higher contrast and more variation in the map and, therefore, more grid points with high density and also high gradient. The resolution dependence and the behavior of the 2D histogram are more obvious when examining the two one-dimensional histograms, where the 2D histogram has been projected along either the gradient or density (Figs. 1*b* and 1*c*). Both the one-dimensional density and gradient histograms showed large variation with resolution. The density histogram at 4.0 Å is close to a Gaussian distribution with only slight skewing at the high-density side. The Gaussian component of the density histogram is attributed to the random overlap of atoms in the map, whereas the skewing is due to the non-overlapping part of the map (Main, 1990). As resolution increases, the density histogram (Fig. 1*b*) becomes more skewed, reflecting the fact that more features in the atoms are resolved and higher peaks in the map are observed. This gives rise to the long tail at the high-density side of the density histogram. The gradient histogram (Fig. 1*c*) at low resolution is also close to a symmetrical distribution. It becomes more skewed with more grid points having high-gradient values as resolution increases, reflecting the fact that more variations are observed in the map at higher resolution. Correlation coefficient and residual, as shown in Tables 1(*a*) and 1(*b*), respectively, can quantify the changes of the 2D histogram with resolution. The average correlation coefficient and residual are 0.735 and 0.257, respectively, for all the resolution steps. The above figures and tables have demonstrated that resolution is a parameter which must be considered when predicting the 2D histograms. In order to eliminate the resolution dependence of the 2D histograms, all the subsequent investigations of the 2D histogram were at a specific resolution while varying other parameters, such as overall *B* factor, structure conformation and phase error.

3.2. *The overall temperature-factor dependence*

The temperature factor reflects the thermal motion of the atoms in the crystal and, therefore, represents the distribution of electrons around the equilibrium position of the atom. The atomic thermal motion will affect the 2D histogram because it changes the distribution of electrons. Here, an example of FGF at different values of overall *B* factor is presented. Fig. 2(*a*) shows six 2D histograms of FGF at 1.6 Å with overall *B* factors at 0, 10, 20, 40, 60 and 80 Å^2. Figs. 2(*b*) and 2(*c*) show the one-dimensional density and gradient histograms for the same conditions. The correlation coefficients and residuals for these 2D histograms of FGF with different overall *B* factors are listed in Tables 2(*a*) and 2(*b*), respectively. From the presented results, it is obvious that the 2D histogram at a given resolution is dependent on the overall *B* factor. The dependence on the overall *B* factor is similar to that of the resolution. As the overall *B* factor increases, the peak in the 2D histogram (Fig. 2*a*) increases, which is correlated with the decrease of the other areas in the 2D histogram. The increase of the overall *B* factor diffuses the electrons around the atom and decreases the peak height at the atomic position in the electron-density map. This decreases the probability of high-density values in the map and causes the

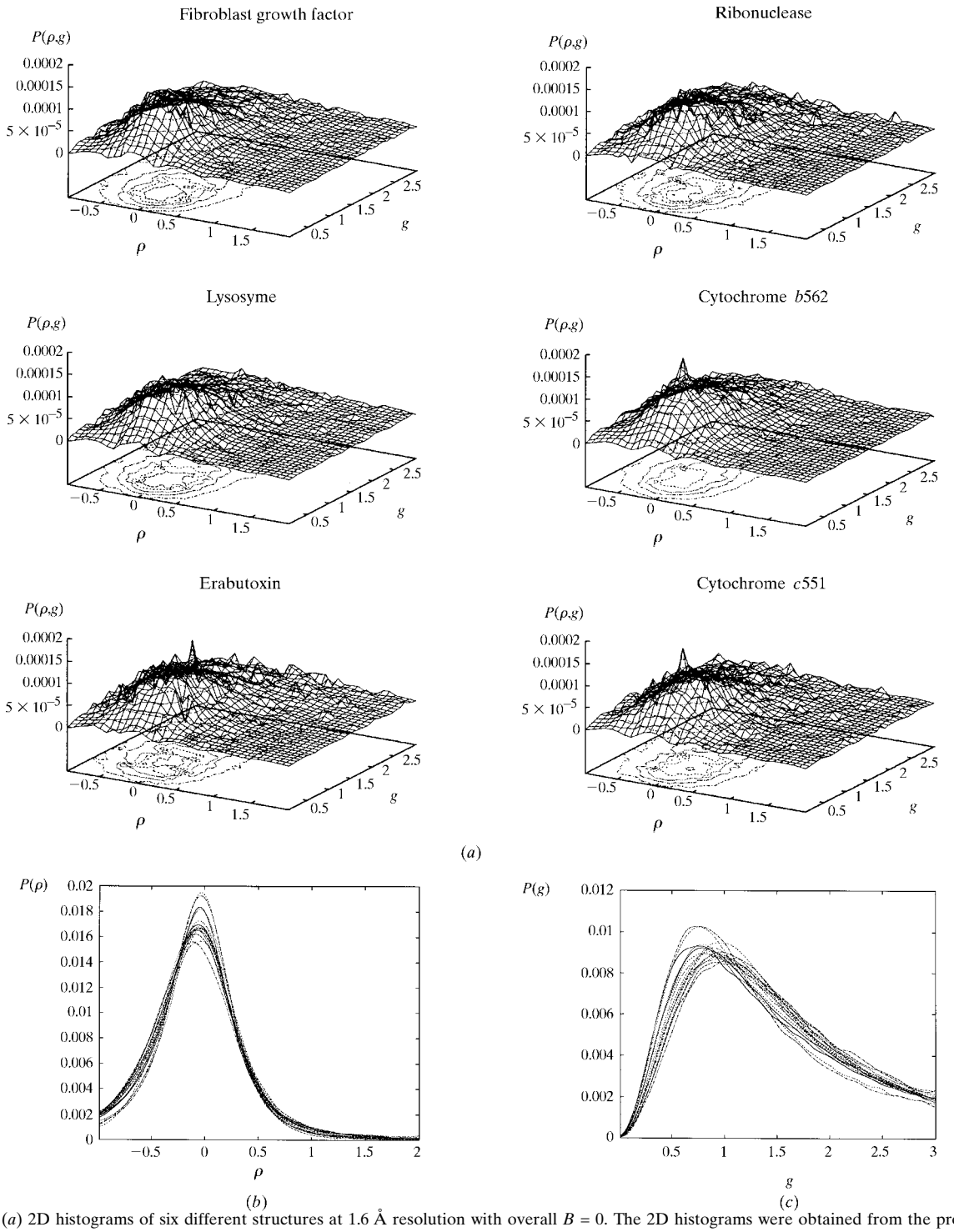

Fig. 3. (*a*) 2D histograms of six different structures at 1.6 Å resolution with overall $B = 0$. The 2D histograms were obtained from the protein region of the maps calculated from the atomic coordinates. The 2D histograms of these six structures of distinctive folds are very similar, revealing the conformation independence of the 2D histogram. Those spikes on the surfaces are due to statistical fluctuations as a result of limited sampling. (*b*) One-dimensional density histograms of 16 different structures at 1.6 Å with $B = 0$. These were obtained from the 2D histograms by integrating over the gradient. The density histograms from these 16 structures are very similar despite that fact the all these structures are from distinctive fold families. The similarity is more pronounced at the high-density side. It is these high densities that contain most of the structural information. (*c*) One-dimensional gradient histograms of 16 different structures at 1.6 Å with $B = 0$. These were obtained from the 2D histograms by integrating over the density. The gradient histograms from these 16 structures of different fold families are very similar.

Table 3. *The 16 representative structures used as a test set*

Structure name, PDB code and folding class according to Orengo *et al.* (1993) are listed.

1. Fibroblast growth factor (4 FGF)	β trefoil	9. Papain (9PAP)	Multidomain
2. Ribonuclease T1 (9RNT)	$\alpha + \beta$ sandwich	10. Ovomucoid (2OVO)	$\alpha + \beta$ S–S rich
3. Lysozyme (1LZ3)	$\alpha + \beta$ mainly α	11. DNA binding protein (2WRP)	α orthogonal
4. Cytochrome B562 (256Ba)	α up/down	12. Cytochrome *b*5 (3B5C)	$\alpha + \beta$ metal rich
5. Erabutoxin (3EBX)	β disulfide rich	13. Endothial aspartic protease (2ER7)	β sandwich
6. Cytochrome *c*551 (451C)	α metal rich	14. Beta trypsin (4PTP)	β Greek key
7. Ca^{2+} binding parvalbumin (4CPV)	α EF-hand	15. Carboxypeptidase A (5CPA)	α/β doubly wound
8. Erythrocruorin (1ECD)	α globin	16. Arabinose-binding protein (8ABP)	α/β multidomain

increase in the probability of the medium-density values (Fig. 2*b*). The increase of the atomic thermal motion also flattens the peak in the map and causes the decrease in the high-gradient values and the accumulation of the low-gradient values in the histogram (Fig. 2*c*). It appears that the effect of the increase of overall *B* factor on the 2D histogram is equivalent to that of the decrease in resolution. Hence, before we examine the structure dependence of the 2D histogram, it should be made independent of the overall *B* factor. This can be achieved by the removal of the overall *B* factor from the structure factors and thereby the map, which corresponds to diffraction from stationary atoms. This standardization of the 2D histogram will not affect the potential implementation of the 2D histogram-matching method, since the overall *B* factor for an unknown structure can be estimated from the observed structure-factor amplitudes using Wilson statistics (Wilson, 1949).

3.3. *The structure dependence*

To examine the dependence of the 2D histogram on structural conformation, we compared 16 protein structures from the Protein Data Bank (PDB). In order to represent different structure types, we selected one structure from each folding class according to the structure classification of Orengo *et al.* (1993). In this way, the representation space is maximized and the chances of local structure dependence of the 2D histograms are minimized. The 16 structures, their PDB codes and their fold families are listed in Table 3.

For each set of comparisons, the histograms were determined at the same resolution and the overall *B* factor was removed, giving a map corresponding to stationary atoms. The 2D histograms of these structures were examined at resolutions ranging from 1.6 to 4.0 Å. The 2D histograms of six representative structures at 1.6 Å are shown in Fig. 3(*a*). The one-dimensional density and gradient histograms of all 16 structures at 1.6 Å are shown in Figs. 3(*b*) and 3(*c*), respectively. The correlation coefficients between each pair of 2D histograms of the 16 structures at three representative resolutions, high (1.6 Å), medium (2.2 Å) and low (3.5 Å), are shown in Tables 4(*a*), 4(*c*) and 4(*e*), respectively. The residuals between the 2D histograms of different structures at the same resolutions are shown in Tables 4(*b*), 4(*d*) and 4(*f*), respectively.

The 2D histograms seem to be very similar among the six structures shown in Fig. 3(*a*). The variation in the 2D histograms between structures of different conformation is far less than between maps of different resolution and also between maps of different *B* factors. Most importantly, as we will show later in §3.4, the conformation dependence of the 2D histogram is far less than the phase-error dependence. The one-dimensional projections show more clearly the conservation of both the density and gradient histograms among various structures. The overall average and variance of the correlation coefficients are 0.904 and 0.037 for the three resolutions examined. The overall average and variance of the residual are 0.132 and 0.029. Whether the amount of variation between the 2D histograms of different structures is significant will be examined in the next section by comparison with the effect of the phase error on the 2D histograms.

3.4. *The phase dependence*

In order for a constraint to be useful for phasing, it must be sensitive to phase errors. We will address the issue of the phase-error sensitivity of the 2D histogram in the following section. To study the sensitivity of the 2D histogram to phase error, we have again selected FGF as a test case. A map was first calculated at 1.6 Å resolution using the correct phases from the atomic coordinates. The overall *B* factor was also removed from the structure-factor amplitudes. Various random phase errors, from 0 to 90° in 10° increments, were then applied to the correct phases. Density and gradient maps were calculated for the structure at each phase error and the 2D histograms were accumulated. The resulting two-dimensional density and gradient histograms with 0, 40 and 90° phase errors are shown in Fig. 4(*a*). The one-dimensional density and one-dimensional gradient histograms with phase errors ranging from 0 to 90° in 10° increments are shown in Figs. 4(*b*) and 4(*c*), respectively. The correlation coefficients and residuals between the 2D histograms are listed in Tables 5(*a*) and 5(*b*).

The results show that the 2D histogram is indeed sensitive to the phase errors. The average correlation coefficient for a 10° phase error is 0.714. This means that

the variation of 2D histograms among structures of drastically different conformation (with a mean correlation of 0.904) is significantly smaller than that caused by a 10° phase error (with an average correlation of 0.714). This establishes the sensitivity of the 2D histogram to phase error. The average correlation coefficients and residuals corresponding to different phase errors are shown in Figs. 5(*a*) and 5(*b*), respectively. The average correlation coefficient for a given phase error was derived from summation of the off-diagonal elements in Table 5(*a*) which corresponds to pairs of 2D histograms with the same phase differences. The average residuals were calculated in the same way from Table 5(*b*). These are marked by diamonds in Figs. 5(*a*) and 5(*b*). A curve of a power series was fitted to those data points and is shown in Figs. 5(*a*) and 5(*b*).

The phase error corresponding to the average correlation coefficient of 0.904 from the 16 structures at three

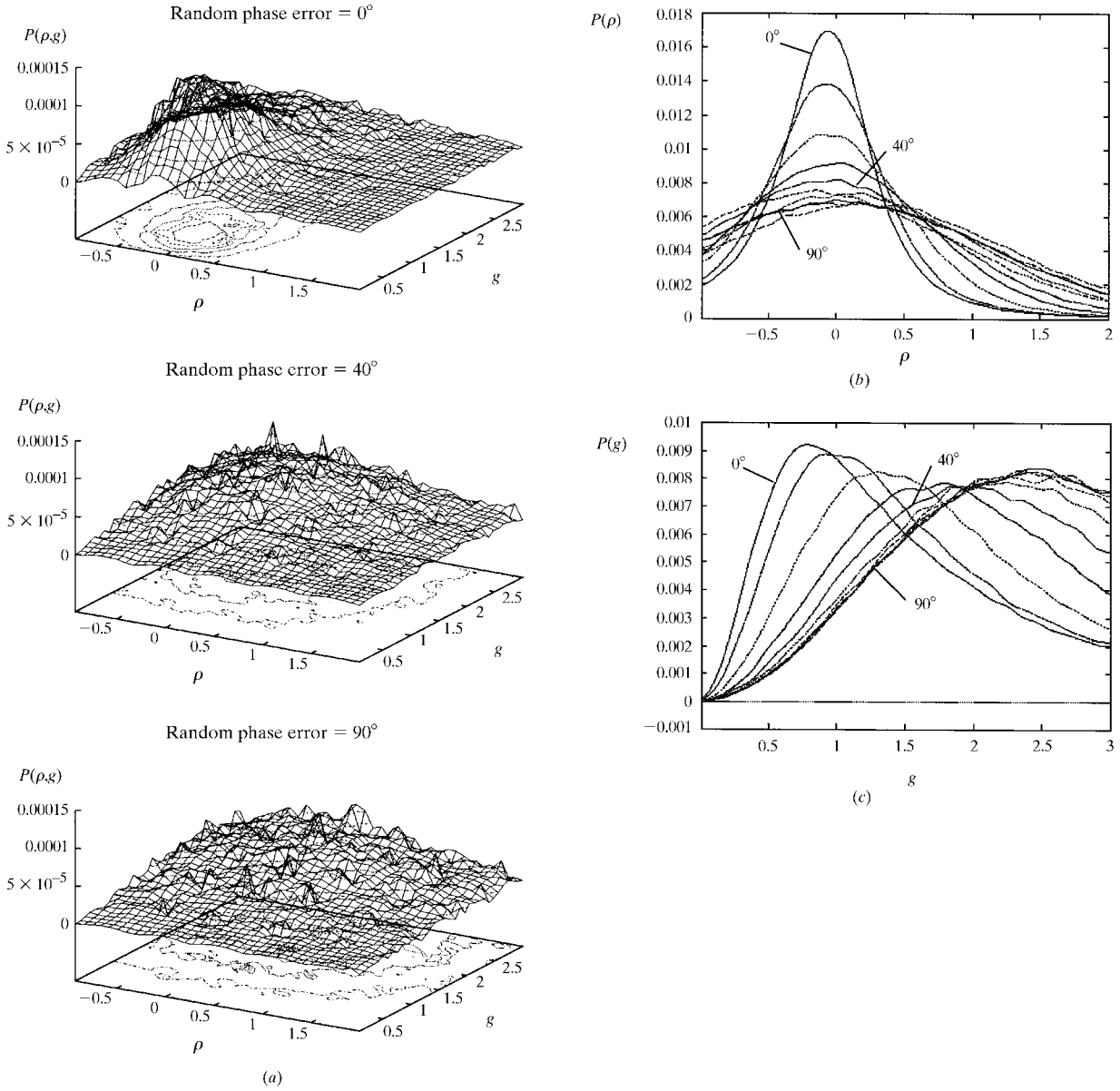

Fig. 4. (*a*) 2D histogram of fibroblast growth factor at 1.6 Å with 0, 40 and 90° phase errors. These 2D histograms differ significantly showing that the 2D histogram is very sensitive to phase error. (*b*) One-dimensional density histogram of fibroblast growth factor at 1.6 Å with phase errors from 0 to 90° in a 10° step. There is a large variation between these ten density histograms showing that the density histogram is dependent on phase error. (*c*) One-dimensional gradient histogram of fibroblast growth factor at 1.6 Å with phase errors from 0 to 90° in 10° steps. All ten gradient histograms differ significantly. This shows that the phase error affects not only the density distribution but also the gradient distribution.

Table 4. *Correlation and residual of 2D histograms between different structures at 1.6, 2.2 and 3.5 Å*

(*a*) Correlation at 1.6 Å. Mean = 0.883, variance = 0.042.

File	4fgf	9rnt	1lz3	256ba	3ebx	451c	4cpv	1ecd	9pap	2ovo	2wrp	3b5c	2er7	4ptp	5cpa	8abp
4fgf	1.000	0.860	0.904	0.906	0.816	0.863	0.903	0.921	0.924	0.876	0.911	0.876	0.927	0.923	0.926	0.929
9rnt	—	1.000	0.858	0.882	0.801	0.827	0.828	0.898	0.862	0.813	0.831	0.843	0.896	0.877	0.884	0.848
1lz3	—	—	1.000	0.906	0.819	0.859	0.883	0.924	0.905	0.858	0.886	0.873	0.928	0.914	0.920	0.901
256ba	—	—	—	1.000	0.850	0.871	0.872	0.951	0.906	0.852	0.873	0.891	0.948	0.926	0.934	0.891
3ebx	—	—	—	—	1.000	0.788	0.779	0.864	0.815	0.764	0.781	0.808	0.862	0.836	0.846	0.800
451c	—	—	—	—	—	1.000	0.840	0.888	0.868	0.820	0.847	0.838	0.893	0.874	0.882	0.857
4cpv	—	—	—	—	—	—	1.000	0.887	0.909	0.869	0.911	0.849	0.900	0.903	0.901	0.923
1ecd	—	—	—	—	—	—	—	1.000	0.923	0.867	0.888	0.906	0.968	0.942	0.953	0.903
9pap	—	—	—	—	—	—	—	—	1.000	0.879	0.916	0.876	0.925	0.928	0.930	0.936
2ovo	—	—	—	—	—	—	—	—	—	1.000	0.875	0.827	0.881	0.878	0.877	0.889
2wrp	—	—	—	—	—	—	—	—	—	—	1.000	0.853	0.904	0.907	0.904	0.933
3b5c	—	—	—	—	—	—	—	—	—	—	—	1.000	0.906	0.889	0.896	0.866
2er7	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.941	0.953	0.909
4ptp	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.939	0.926
5cpa	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.921
8abp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000

(*b*) Residual at 1.6 Å. Mean = 0.145, variance = 0.027.

File	4fgf	9rnt	1lz3	256ba	3ebx	451c	4cpv	1ecd	9pap	2ovo	2wrp	3b5c	2er7	4ptp	5cpa	8abp
4fgf	0.000	0.157	0.133	0.126	0.181	0.159	0.146	0.116	0.119	0.162	0.141	0.150	0.113	0.120	0.115	0.115
9rnt	—	0.000	0.159	0.143	0.188	0.177	0.182	0.133	0.154	0.190	0.181	0.167	0.139	0.147	0.142	0.160
1lz3	—	—	0.000	0.128	0.180	0.162	0.157	0.115	0.131	0.170	0.153	0.152	0.116	0.126	0.120	0.131
256ba	—	—	—	0.000	0.164	0.152	0.156	0.090	0.123	0.169	0.155	0.138	0.098	0.111	0.103	0.129
3ebx	—	—	—	—	0.000	0.196	0.205	0.156	0.180	0.210	0.205	0.187	0.162	0.171	0.166	0.185
451c	—	—	—	—	—	0.000	0.180	0.143	0.154	0.190	0.178	0.173	0.144	0.153	0.147	0.159
4cpv	—	—	—	—	—	—	0.000	0.146	0.145	0.170	0.142	0.171	0.139	0.143	0.144	0.133
1ecd	—	—	—	—	—	—	—	0.000	0.111	0.160	0.146	0.129	0.080	0.098	0.088	0.118
9pap	—	—	—	—	—	—	—	—	0.000	0.161	0.140	0.149	0.112	0.115	0.109	0.110
2ovo	—	—	—	—	—	—	—	—	—	0.000	0.167	0.183	0.155	0.159	0.158	0.154
2wrp	—	—	—	—	—	—	—	—	—	—	0.000	0.171	0.136	0.141	0.141	0.125
3b5c	—	—	—	—	—	—	—	—	—	—	—	0.000	0.134	0.140	0.136	0.152
2er7	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.104	0.093	0.114
4ptp	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.103	0.114
5cpa	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.113
8abp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000

(*c*) Correlation at 2.2 Å. Mean = 0.892, variance = 0.041.

File	4fgf	9rnt	1lz3	256ba	3ebx	451c	4cpv	1ecd	9pap	2ovo	2wrp	3b5c	2er7	4ptp	5cpa	8abp
4fgf	1.000	0.831	0.887	0.899	0.795	0.847	0.863	0.927	0.908	0.838	0.850	0.852	0.916	0.908	0.914	0.908
9rnt	—	1.000	0.823	0.855	0.764	0.789	0.783	0.866	0.835	0.762	0.738	0.809	0.859	0.844	0.852	0.826
1lz3	—	—	1.000	0.898	0.789	0.838	0.861	0.923	0.897	0.831	0.837	0.851	0.913	0.902	0.908	0.900
256ba	—	—	—	1.000	0.810	0.850	0.869	0.944	0.906	0.839	0.827	0.876	0.931	0.919	0.929	0.908
3ebx	—	—	—	—	1.000	0.754	0.739	0.823	0.796	0.719	0.695	0.770	0.821	0.801	0.804	0.784
451c	—	—	—	—	—	1.000	0.815	0.875	0.855	0.789	0.791	0.804	0.874	0.854	0.859	0.853
4cpv	—	—	—	—	—	—	1.000	0.901	0.886	0.858	0.889	0.820	0.901	0.879	0.895	0.908
1ecd	—	—	—	—	—	—	—	1.000	0.937	0.868	0.874	0.891	0.954	0.945	0.953	0.939
9pap	—	—	—	—	—	—	—	—	1.000	0.859	0.879	0.860	0.924	0.919	0.925	0.928
2ovo	—	—	—	—	—	—	—	—	—	1.000	0.857	0.791	0.872	0.848	0.863	0.878
2wrp	—	—	—	—	—	—	—	—	—	—	1.000	0.784	0.857	0.862	0.875	0.910
3b5c	—	—	—	—	—	—	—	—	—	—	—	1.000	0.883	0.869	0.877	0.860
2er7	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.924	0.933	0.922
4ptp	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.933	0.923
5cpa	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.934
8abp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000

different resolutions is marked by the vertical dashed line in Fig. 5(*a*). This extrapolation gives a phase error of about 2°. The phase error corresponding to the average residual of 0.132 from the 16 structures at three different resolutions is marked by the vertical dashed line in Fig. 5(*b*). This shows a phase error of about 5°. The difference between the two extrapolated phase-error values reflects the sensitivity difference between the correlation coefficient and the residual over this region of phase error. The combined estimate of the phase

Table 4 (*cont.*)

(*d*) Residual at 2.2 Å. Mean = 0.136, variance = 0.030.

File	4fgf	9rnt	1lz3	256ba	3ebx	451c	4cpv	1ecd	9pap	2ovo	2wrp	3b5c	2er7	4ptp	5cpa	8abp
4fgf	0.000	0.160	0.136	0.126	0.179	0.159	0.164	0.109	0.120	0.178	0.169	0.155	0.122	0.123	0.118	0.120
9rnt	—	0.000	0.165	0.147	0.192	0.184	0.193	0.143	0.155	0.204	0.210	0.171	0.153	0.154	0.148	0.157
1lz3	—	—	0.000	0.129	0.183	0.166	0.165	0.114	0.130	0.180	0.174	0.157	0.126	0.128	0.123	0.127
256ba	—	—	—	0.000	0.173	0.157	0.155	0.094	0.119	0.173	0.172	0.141	0.112	0.112	0.104	0.115
3ebx	—	—	—	—	0.000	0.200	0.214	0.168	0.177	0.223	0.227	0.193	0.174	0.177	0.174	0.180
451c	—	—	—	—	—	0.000	0.186	0.145	0.154	0.199	0.197	0.180	0.149	0.157	0.153	0.154
4cpv	—	—	—	—	—	—	0.000	0.140	0.153	0.173	0.160	0.180	0.134	0.155	0.147	0.141
1ecd	—	—	—	—	—	—	—	0.000	0.100	0.160	0.151	0.133	0.095	0.095	0.087	0.095
9pap	—	—	—	—	—	—	—	—	0.000	0.169	0.158	0.149	0.113	0.114	0.110	0.107
2ovo	—	—	—	—	—	—	—	—	—	0.000	0.179	0.195	0.154	0.172	0.166	0.158
2wrp	—	—	—	—	—	—	—	—	—	—	0.000	0.195	0.143	0.146	0.140	0.147
2er7	—	—	—	—	—	—	—	—	—	—	—	0.000	0.154	0.163	0.158	0.144
3b5c	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.118	0.110	0.110
4ptp	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.104	0.111
5cpa	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.103
8abp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000

(*e*) Correlation at 3.5 Å. Mean = 0.938, variance = 0.029.

File	4fgf	9rnt	1lz3	256ba	3ebx	451c	4cpv	1ecd	9pap	2ovo	2wrp	3b5c	2er7	4ptp	5cpa	8abp
4fgf	1.000	0.935	0.964	0.962	0.917	0.946	0.920	0.973	0.967	0.923	0.914	0.947	0.953	0.970	0.958	0.972
9rnt	—	1.000	0.935	0.948	0.914	0.937	0.909	0.947	0.957	0.911	0.868	0.940	0.941	0.938	0.930	0.953
1lz3	—	—	1.000	0.966	0.912	0.941	0.908	0.973	0.963	0.914	0.887	0.947	0.946	0.966	0.962	0.967
245ba	—	—	—	1.000	0.914	0.950	0.919	0.973	0.972	0.927	0.869	0.960	0.959	0.967	0.974	0.972
3ebx	—	—	—	—	1.000	0.911	0.862	0.923	0.928	0.866	0.849	0.905	0.906	0.908	0.889	0.923
451c	—	—	—	—	—	1.000	0.918	0.951	0.958	0.918	0.886	0.942	0.949	0.946	0.939	0.957
4cpv	—	—	—	—	—	—	1.000	0.925	0.936	0.950	0.921	0.925	0.962	0.933	0.928	0.945
1ecd	—	—	—	—	—	—	—	1.000	0.975	0.925	0.911	0.957	0.955	0.980	0.970	0.981
9pap	—	—	—	—	—	—	—	—	1.000	0.937	0.904	0.961	0.969	0.968	0.961	0.978
2ovo	—	—	—	—	—	—	—	—	—	1.000	0.897	0.926	0.961	0.932	0.936	0.943
2wrp	—	—	—	—	—	—	—	—	—	—	1.000	0.881	0.906	0.922	0.883	0.923
3b5c	—	—	—	—	—	—	—	—	—	—	—	1.000	0.953	0.953	0.953	0.963
2er7	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.955	0.958	0.968
4ptp	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.970	0.978
5cpa	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000	0.970
8abp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.000

(*f*) Residual at 3.5 Å. Mean = 0.114, variance = 0.029.

File	4fgf	9rnt	1lz3	256ba	3ebx	451c	4cpv	1ecd	9pap	2ovo	2wrp	3b5c	2er7	4ptp	5cpa	8abp
4fgf	0.000	0.110	0.087	0.086	0.129	0.106	0.160	0.076	0.084	0.146	0.124	0.106	0.114	0.080	0.097	0.079
9rnt	—	0.000	0.110	0.099	0.137	0.114	0.156	0.099	0.093	0.148	0.146	0.110	0.121	0.106	0.114	0.095
1lz3	—	—	0.000	0.084	0.135	0.109	0.164	0.076	0.087	0.149	0.137	0.106	0.118	0.084	0.093	0.083
256ba	—	—	—	0.000	0.136	0.100	0.149	0.070	0.074	0.134	0.136	0.091	0.104	0.077	0.075	0.071
3ebx	—	—	—	—	0.000	0.140	0.207	0.132	0.129	0.193	0.167	0.148	0.161	0.141	0.160	0.133
451c	—	—	—	—	—	0.000	0.153	0.100	0.094	0.143	0.141	0.111	0.114	0.104	0.112	0.095
4cpv	—	—	—	—	—	—	0.000	0.151	0.137	0.105	0.145	0.140	0.103	0.140	0.130	0.129
1ecd	—	—	—	—	—	—	—	0.000	0.070	0.139	0.123	0.094	0.107	0.064	0.078	0.063
9pap	—	—	—	—	—	—	—	—	0.000	0.127	0.124	0.090	0.092	0.078	0.087	0.067
2ovo	—	—	—	—	—	—	—	—	—	0.000	0.151	0.131	0.098	0.131	0.118	0.121
2wrp	—	—	—	—	—	—	—	—	—	—	0.000	0.143	0.125	0.120	0.140	0.116
3b5c	—	—	—	—	—	—	—	—	—	—	—	0.000	0.107	0.096	0.095	0.088
2er7	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.106	0.096	0.092
4ptp	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.077	0.068
5cpa	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000	0.077
8abp	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.000

error from both the correlation coefficient and the residual is about 4°. This establishes the minimum phase error that the ideal 2D histogram, derived from consensus 2D histograms of different structures, could specify when used as a target for density modification or phase discrimination. Note that structures are solved using MIR phases with phase errors as high as 70°. Structures can be solved routinely using MIR phases with errors below 50°. Here, only random errors from the MIR phases are compared, since the 2D histogram is

a statistically derived constraint. When there are systematic errors in the MIR phases, the distortion to the map is more severe and maps can only be interpreted with much smaller phase errors. Because the variation between the 2D histograms of different structures corresponds to that caused by a very small phase error, we conclude that the 2D histogram is independent of structure, with respect to a tolerable amount of phase error. This established the predictability of the 2D histogram for unknown structures. The ideal 2D histogram for an unknown structure can be taken as the consensus 2D histogram of a representative set of well refined structures as listed in Table 3. The sensitivity of the 2D histogram to phase error suggests that the 2D histogram can be used as a constraint for phase refinement and extension. It is also anticipated that the 2D histograms could be used as a figure of merit to assess the accuracy of phase sets in an *ab initio* phasing approach.

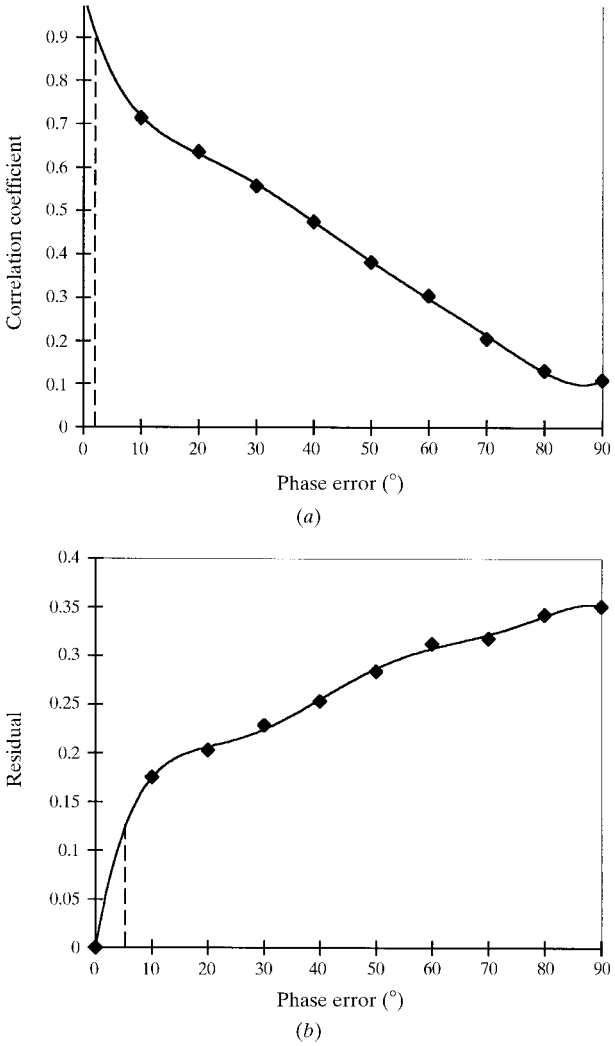

Fig. 5. (*a*) Correlation coefficients of 2D histograms as a function of phase errors. The correlation coefficients are derived from Table 5(*a*) by averaging over all the pairs of 2D histograms with the same phase difference. The correlation coefficients are measured between 0° and 90° with a 10° interval and are represented as diamonds in the figure. A curve of a power series is fitted with the measured correlation coefficients. The vertical line indicates the average correlation coefficients between different structures and the corresponding phase error. (*b*) Residual of 2D histograms as a function of phase errors. The residuals are derived from Table 5(*b*) by averaging over all the pairs of 2D histograms with the same phase difference. The residuals are measured between 0 and 90° with a 10° interval and are represented as diamonds in the figure. A curve of a power series is fitted with the measured residuals. The vertical line indicates the average residual between different structures and the corresponding phase error.

4. Discussion

We have demonstrated in §3 that the 2D histogram is independent of structure and dependent on resolution, overall *B* factor and phase error. We can standardize 2D histograms by removing the overall *B* factor from the electron-density map, since the overall *B* factor can be estimated for unknown structures from Wilson statistics (Wilson, 1949). The resolution dependence of the 2D histogram can be eliminated by the use of resolution-specific 2D histograms. By making the constraints resolution specific, the non-atomic resolution diffraction of protein crystals can be more effectively dealt with. It also offers an advantage over other resolution-independent constraints on phasing extension, since the constraint is specified for each resolution.

By standardizing the 2D histogram with the removal of the overall *B* factor and making it resolution specific, a 2D histogram is only dependent on the phase error. The ideal 2D histogram for an unknown structure at a given resolution can be predicted by using the consensus 2D histogram derived from known structures. We can systematically adjust the electron-density values for a given map so that the modified map will have the same 2D histogram as the ideal one. This 2D histogram-matching procedure will greatly improve the quality of the map and, therefore, the phases.

The independence of molecular conformation and sensitivity to phase error has also been examined (Xiang & Carter, 1996) on one-dimensional density, gradient and Laplacian histograms and 2D histograms of the pairwise combinations of these three components. The one-dimensional histograms of density, gradient and Laplacian derived from model structures of an α-helix, β-strand and loop were compared and it was found that these one-dimensional histograms were independent of molecular conformations. The sensitivity of the one-dimensional and 2D histograms to phase errors was also investigated and the gradient histogram was found to be the most sensitive to phase error. Our study focused on the 2D histogram of density and gradient and extended the scope of examination by including protein structures from 16 different fold families instead of model struc-

Table 5. *Correlation and residual of 2D histograms of fibroblast growth factor at 1.6 Å with phase errors from 0 to 90°*

(*a*) Correlation. Mean = 0.497, variance = 0.215.

Phase error (°)	0	10	20	30	40	50	60	70	80	90
0	1.000	0.868	0.670	0.433	0.285	0.190	0.116	0.065	0.076	0.109
10	—	1.000	0.770	0.565	0.411	0.291	0.211	0.140	0.217	0.188
20	—	—	1.000	0.732	0.601	0.469	0.435	0.363	0.379	0.334
30	—	—	—	1.000	0.709	0.621	0.582	0.595	0.545	0.583
40	—	—	—	—	1.000	0.686	0.634	0.663	0.590	0.599
50	—	—	—	—	—	1.000	0.665	0.665	0.662	0.656
60	—	—	—	—	—	—	1.000	0.680	0.683	0.682
70	—	—	—	—	—	—	—	1.000	0.655	0.649
80	—	—	—	—	—	—	—	—	1.000	0.663
90	—	—	—	—	—	—	—	—	—	1.000

(*b*) Residual. Mean = 0.244, variance = 0.068.

Phase error (°)	0	10	20	30	40	50	60	70	80	90
0	0.000	0.144	0.207	0.269	0.306	0.329	0.348	0.360	0.357	0.351
10	—	0.000	0.173	0.233	0.273	0.303	0.321	0.337	0.321	0.327
20	—	—	0.000	0.169	0.207	0.241	0.253	0.267	0.264	0.274
30	—	—	—	0.000	0.248	0.283	0.299	0.297	0.313	0.301
40	—	—	—	—	0.000	0.169	0.181	0.177	0.192	0.189
50	—	—	—	—	—	0.000	0.168	0.169	0.169	0.169
60	—	—	—	—	—	—	0.000	0.171	0.171	0.171
70	—	—	—	—	—	—	—	0.000	0.169	0.173
80	—	—	—	—	—	—	—	—	0.000	0.168
90	—	—	—	—	—	—	—	—	—	0.000

tures. Our studies based on a set of complete protein structures have reached the same conclusion. Moreover, by quantifying the correlation and residual between all the 2D histograms examined, we have derived the variation of 2D histograms among different structures and established the corresponding phase error as the upper limit for any density-modification method that uses the 2D histogram as the constraint.

The application of the local environment of the electron density as a constraint for phase improvement has been demonstrated by Refaat *et al.* (1996). They proposed a method of density modification based on density histograms that takes into account the local environment of electron density. The characteristics investigated are the local minimum, maximum and variance of the density. Tests of this method on 2Zn insulin and RNAp1 structures have shown it to reduce the phase errors by a further 10° compared with the normal histogram-matching method.

Our investigation and the above studies strongly suggest that the information about the local environment of electron density, such as the local minimum, maximum and mean density, could be exploited for phase improvement. The 2D histogram of the density and gradient is another measure of the characteristic features of the density and its local environment. The 2D histogram substantially reduces the degeneracy of the one-dimensional electron-density histogram, thereby providing a more accurate target for phase improvement and a more discriminating figure of merit for detecting phase errors. A systematic approach of examining the resolution, temperature-factor, conformation and phase dependence of a given constraint has been adopted in our study of 2D histograms. This approach can be generalized to investigate the multi-dimensional probability distribution of electron density, such as the Laplacian (Xiang & Carter, 1996) and also the higher order derivatives of electron density. A consensus 2D histogram can be derived from the 16 structures and can be used as the standard for 2D histogram matching. Methods for the matching of the 2D histograms of a given map to that of the standard one are being developed and tested. The 2D histogram matching will be used to improve the quality of the map and, therefore, the accuracy of the phases. This density-modification technique will be used not only to refine but also to extend the phases to higher resolution. Using the 2D histogram as a constraint instead of the one-dimensional histogram will provide more phasing power and, therefore, the 2D histogram-matching method could potentially be more powerful in phasing refinement and extension.

As mentioned previously, the density histogram discards all positional information. Although the histogram is unique for any particular map, vastly different maps can have identical histograms. We have attempted to remedy this problem by providing further constraints, such as the probability distribution of the gradient, in addition to the electron-density distribution. It is foreseeable that these constraints might still not be capable of giving a unique solution to the electron-density equation. Other constraints of different characteristics could further reduce the degeneracy of the constraint space. Since the electron-density histogram was found to

be synergistic with other constraints like solvent flatness, equal molecules, atomic shape and map continuity (Zhang, 1993; Zhang & Main, 1990*b*), the synergism between the 2D histogram and the above constraints will be investigated.

We thank Dr Peter Main and Dr David Baker for discussion. This work was supported by funds from the Fred Hutchinson Cancer Research Center and the National Institutes of Health grant R29GM55663.

References

Arnold, E. & Rossmann, M. G. (1986). *Proc. Natl Acad. Sci. USA*, **83**, 5489–5493.
Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 429–439.
Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 186–192.
Bricogne, G. (1976). *Acta Cryst.* A**32**, 832–847.
Bystroff, C., Baker, D., Fletterick, R. J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 440–448.
Harrison, R. W. (1988). *J. Appl. Cryst.* **21**, 949–952.
Hauptman, H. (1986). *Science*, **233**, 178–183.
Karle, J. (1986). *Science*, **232**, 837–843.
Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 134–136.
Lunin, V. Y. (1988). *Acta Cryst.* A**44**, 144–150.
Main, P. (1990). *Acta Cryst.* A**46**, 507–509.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Matthews, B. W. (1974). *J. Mol. Biol.* **82**, 513–526.
Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). *Protein Eng.* **6**(5), 485–500.
Refaat, L. S., Tate, C. & Woolfson, M. M. (1996). *Acta Cryst.* D**52**, 252–256.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–113.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Wilson, C. & Agard, D. A. (1993). *Acta Cryst.* A**49**, 97–104.
Xiang, S. & Carter, C. W. Jr (1996). *Acta Cryst.* D**52**, 49–56.
Zhang, K. Y. J. (1993). *Acta Cryst.* D**49**, 213–222.
Zhang, K. Y. J., Cowtan, K. D. & Main, P. (1997). *Methods Enzymol.* **277**, 53–64.
Zhang, K. Y. J. & Main, P. (1990*a*). *Acta Cryst.* A**46**, 41–46.
Zhang, K. Y. J. & Main, P. (1990*b*). *Acta Cryst.* A**46**, 377–381.