Acta Cryst. (1996). D**52**, 49–56

Representing Stereochemical Information in Macromolecular Electron-Density Distributions by Multi-dimensional Histograms

SHIBIN XIANG* AND CHARLES W. CARTER JR

Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC27599-7260, USA

(Received 2 March 1995; accepted 4 July 1995)

Abstract

Previous studies have demonstrated the value of ideal electron-density histograms as targets for the corresponding histograms of experimental electron-density maps. The electron-density histogram makes use of density values as independent objects, and no relationship between them is taken into account. Extension to include the relationships between neighboring density values leads naturally to a multi-dimensional histogram defined as the joint frequency of the density values and their higher order derivatives. We show here that the multi-dimensional histogram including additional dimensions composed of the gradient magnitude and Laplacian of the density is minimally dependent on molecular folding and packing, and captures substantially more stereochemical information than the conventional electron-density histogram. The gradient histogram appears to be much more sensitive to phase errors than the conventional electron-density histogram. Potential uses of the multi-dimensional histogram include improved targets for density modification and more reliable figures of merit for evaluating correct phases.

1. Introduction

Prior information about the electron-density distribution provides a link between the amplitudes and phases of structure factors on which direct methods are founded (Bricogne, 1984; Hauptman, 1986; Karle, 1986; Woolfson, 1987). The most widely used information has been the non-negativity of electron density, as explicitly employed in the derivation of inequality relationships between structure factors (Harker & Kasper, 1948; Karle & Hauptman, 1950) and implicitly in maximum-entropy methods (Collins, 1978; Bricogne, 1984, 1988). The more restrictive constraint that $\int \rho^3(x)\mathrm{d}v$ is a maximum was used in deriving the Cochran distribution (Cochran, 1952). The local shape or atomicity of molecular electron densities also gives rise to Sayre's equation (Sayre, 1952) which has proven useful in macromolecular phase extension and refinement (Sayre, 1974; Main, 1990; Zhang, 1993).

In various density-modification methods, prior information about the electron-density distribution has been used for map improvement by actually replacing the value of a particular density, ρ, by a 'target' value, ρ_t, obtained from a prior expectation. These methods differ in the choice of the expected values. Examples include filtering the electron density such as $\rho_{\min} < \rho_t(\mathbf{r}) < \rho_{\max}$ (Cannillo, Oberti & Ungaretti, 1983), and restraining the electron density locally to conform to criteria like $\rho_t = 3\rho^2 - 2\rho^3$ (Collins, Brice, La Cour & Legg, 1976).

The ideal electron-density histogram, which specifies not only permitted values for the electron density but also their frequencies (Lunin, 1988; Harrison, 1988; Zhang & Main, 1990*a*) provides a better modification target than those described above because it represents more of the prior chemical information. A distinguishing characteristic of the density histogram is that the probability of a particular density value occurring in the unit cell encodes some stereochemical information. Use of an ideal histogram as a target in macromolecular phase extension and refinement has shown promising results (Zhang & Main, 1990*a*,*b*; Zhang, 1993). A related indicator of the utility of the density histogram is that it also can serve as a figure of merit to retrieve correct phase sets in *ab initio* phasing of macromolecules (Lunin, Urzhumtsev & Skovoroda, 1990).

The electron-density histogram is determined by the characteristic shape of a molecular electron-density distribution, which in turn depends on stereochemical features such as bond lengths and angles between atoms in the molecule. The electron-density histogram is insufficient to represent all these features uniquely, and is degenerate in the sense that many electron-density distributions can be constructed to fit a target histogram without necessarily having the correct molecular shape (Lunin *et al.*, 1990). In this paper, a multi-dimensional histogram is proposed that provides a more comprehensive representation of stereochemical constraints on the electron-density distribution. The multi-dimensional histogram substantially reduces the degeneracy of the electron-density histogram, thereby providing more discriminating targets for density modification and figures of merit for detecting phase errors.

2. The multi-dimensional histogram

Stereochemical information is usually expressed as bond lengths and bond angles between atoms in an atomic model. This information can be easily imposed as restraints on atomic positions during structure refinements (Brünger, 1992; Tronrud, Ten Eyck & Matthews, 1987; Hendrickson, 1985). However, it cannot be applied in this form to the electron-density distribution since the objects in the density distribution are not atomic positions but pixels of electron density. Nevertheless, in macromolecular structure determination, the characteristic geometrical shape of the electron density that provides a unique guide for the crucial step of model building, derives implicitly from the same bond lengths, bond angles and atom types. Thus, this characteristic geometrical shape expresses the stereochemical information.

Here, we explore a representation of this characteristic geometrical shape which can be applied on the electron-density distribution $\rho(\mathbf{r})$ within a molecular envelope and can be easily realized in computer algorithms.

2.1. *Definition*

Geometrical shape at a particular position $\mathbf{r}_0$ in an electron-density distribution is not completely defined by the electron-density value $\rho(\mathbf{r}_0)$. Complementary information is provided by the derivatives $\nabla\rho(\mathbf{r}_0)$, $\nabla^2\rho(\mathbf{r}_0)$, $\nabla^3\rho(\mathbf{r}_0)$, and so forth. Here, ∇ is the gradient operator (Borden, 1983),

$$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}, \qquad (1)$$

and its scalar products are written as $\nabla^{n+1} = \nabla^n \cdot \nabla$. These derivatives are appropriate components to describe the stereochemical information in the electron-density distribution $\rho(\mathbf{r})$. In the conventional electron-density histogram, only the density $\rho(\mathbf{r}_0)$ is used. To capture more stereochemical information, the derivatives should be taken into account.

There is no restriction on the number of components used to construct the multi-dimensional histogram. The more components used, the more stereochemical information can be encoded in the multi-dimensional histogram. However, in consideration of the computational costs and since the low-order derivatives carry most of the information about the characteristic shape of the molecular electron-density distribution, we will confine ourselves to the density $\rho(\mathbf{r}_0)$ and its two lowest order derivatives $\nabla\rho(\mathbf{r}_0)$ and $\nabla^2\rho(\mathbf{r}_0)$. Higher dimensional histograms can be obtained in a similar manner by including more derivatives. Since an appropriate representation of stereochemical information should be independent of molecular orientation, we used the gradient magnitude, $|\nabla\rho(\mathbf{r}_0)|$ to replace the gradient $\nabla\rho(\mathbf{r}_0)$ in constructing the multi-dimensional histogram.

Suppose that $V(\tau_0, \tau_1, \tau_2)$ is the real-space volume taken by the electron densities with $\rho(\mathbf{r}_0) = \tau_0$, $|\nabla\rho(\mathbf{r}_0)| = \tau_1$, $\nabla^2\rho(\mathbf{r}_0) = \tau_2$ and that the values of these scalar functions are divided into bins with the lengths $\Delta\tau_0$, $\Delta\tau_1$ and $\Delta\tau_2$, respectively. We define the three-dimensional histogram, $\mathbf{egl}(\tau_0, \tau_1, \tau_2)$, as the joint frequency of the electron densities found inside the bin $[\tau_0 - (\Delta\tau_0/2),\ \tau_0 + (\Delta\tau_0/2);\ \tau_1 - (\Delta\tau_1/2),\ \tau_1 + (\Delta\tau_1/2);\ \tau_2 - (\Delta\tau_2/2),\ \tau_2 + (\Delta\tau_2/2)]$,

$$\mathbf{egl}(\tau_0, \tau_1, \tau_2) = [\Delta V(\tau_0, \tau_1, \tau_2)]/(\Delta\tau_0\Delta\tau_1\Delta\tau_2). \qquad (2)$$

$\mathbf{egl}(\tau_0, \tau_1, \tau_2)$ defined in (2) is convenient for practical computation since the geometrical properties $\rho(\mathbf{r}_0)$, $|\nabla\rho(\mathbf{r}_0)|$, $\nabla^2\rho(\mathbf{r}_0)$ are always calculated on discrete grid points. However it depends, in the strict sense, on the grid and on the bin lengths used in the calculations. A more accurate form can be obtained by considering the limiting case that the number of grid points increase indefinitely and the bin lengths tend to zero,

$$\mathbf{EGL}(\tau_0, \tau_1, \tau_2) = [\partial^3 V(\tau_0, \tau_1, \tau_2)]/(\partial\tau_0\partial\tau_1\tau_2). \qquad (3)$$

2.2. *Projections*

Several useful one- and two-dimensional histograms can be obtained by projecting the histogram **EGL** onto the corresponding sub-space. Projections of the **EGL** onto corresponding components give rise to the following one-dimensional histograms.

The conventional electron-density histogram, which has been used in macromolecular electron-density modification applications (Zhang & Main, 1990*a*; Zhang, 1993; Lunin, 1993),

$$\mathbf{E}(\tau_0) = \int\int \mathbf{EGL}(\tau_0, \tau_1, \tau_2)\mathrm{d}\tau_1\,\mathrm{d}\tau_2. \qquad (4)$$

Gradient histogram,

$$\mathbf{G}(\tau_1) = \int\int \mathbf{EGL}(\tau_0, \tau_1, \tau_2)\mathrm{d}\tau_0\,\mathrm{d}\tau_2. \qquad (5)$$

Laplacian histogram,

$$\mathbf{L}(\tau_2) = \int\int \mathbf{EGL}(\tau_0, \tau_1, \tau_2)\mathrm{d}\tau_0\,\mathrm{d}\tau_1. \qquad (6)$$

Projections along the same components give rise to the following two-dimensional histograms.

Density-gradient histogram,

$$\mathbf{EG}(\tau_0, \tau_1) = \int \mathbf{EGL}(\tau_0, \tau_1, \tau_2)\,\mathrm{d}\tau_2. \qquad (7)$$

Density–Laplacian histogram,

$$\mathbf{EL}(\tau_0, \tau_2) = \int \mathbf{EGL}(\tau_0, \tau_1, \tau_2)\,\mathrm{d}\tau_1. \qquad (8)$$

Gradient–Laplacian histogram,

$$\mathbf{GL}(\tau_1, \tau_2) = \int \mathbf{EGL}(\tau_0, \tau_1, \tau_2)\,\mathrm{d}\tau_0. \qquad (9)$$

Since $\rho(\mathbf{r}_0)$, $|\nabla\rho(\mathbf{r}_0)|$ and $\nabla^2\rho(\mathbf{r}_0)$ each characterize different properties of the molecular shape, they reflect somewhat independent stereochemical information. Therefore, a multi-dimensional histogram always includes more stereochemical information than do its projections and will, consequently, be more sensitive to phase errors than its projections, as we will demonstrate below.

2.3. *Independence of molecular conformation*

A necessary requirement for an appropriate representation of stereochemical information is that it must be independent of molecular conformation. To show that the histogram **EGL** satisfies this criteria, we rewrite (3) as,

$$\mathbf{EGL}(\tau_0, \tau_1, \tau_2) = (1/\tau_0)[\partial^3 N(\tau_0, \tau_1, \tau_2)]/(\partial\tau_0\partial\tau_1\partial\tau_2), \tag{10a}$$

here $N(\tau_0, \tau_1, \tau_2)$ is the number of electrons in $V(\tau_0, \tau_1, \tau_2)$,

$$N(\tau_0, \tau_1, \tau_2) = \int_{V(\tau_0,\tau_1,\tau_2)} \rho(\mathbf{r})\,d\mathbf{r}. \tag{10b}$$

For an ideal electron-density distribution $\rho(\mathbf{r})$, the molecular volume V_T can be divided, by selecting an appropriate electron-density threshold ρ_a, into two regions (Fig. 1): $V_{\text{nonbond}}(\tau_0, \tau_1, \tau_2) \in \{\tau_0 < \rho_a\}$ and $V_{\text{covalent}}(\tau_0, \tau_1, \tau_2) \in \{\tau_0 \geq \rho_a\}$ such that the latter is occupied by the density, $\rho_{\text{covalent}}(\mathbf{r})$, from main-chain and side groups of the molecule and the former by the density, $\rho_{\text{nonbond}}(\mathbf{r})$, from nearby residues that interact with each other either by van der Waals contacts or hydrogen bonds. Therefore, (10) can be written as,

$$\mathbf{EGL}(\tau_0, \tau_1, \tau_2) = (1/\tau_0)\left\{[\partial^3 N_{\text{nonbond}}(\tau_0, \tau_1, \tau_2)] \div (\partial\tau_0\partial\tau_1\partial\tau_2)\right\}, \quad \tau_0 < \rho_a, \tag{11a}$$

$$\mathbf{EGL}(\tau_0, \tau_1, \tau_2) = (1/\tau_0)\left\{[\partial^3 N_{\text{covalent}}(\tau_0, \tau_1, \tau_2)] \div (\partial\tau_0\partial\tau_1\partial\tau_2)\right\}, \quad \tau_0 \geq \rho_a, \tag{11b}$$

$$N_{\text{nonbond}}(\tau_0, \tau_1, \tau_2) = \int_{V_{\text{nonbond}}(\tau_0,\tau_1,\tau_2)} \rho_{\text{nonbond}}(\mathbf{r})\,d\mathbf{r}, \quad \tau_0 < \rho_a, \tag{11c}$$

$$N_{\text{covalent}}(\tau_0, \tau_1, \tau_2) = \int_{V_{\text{covalent}}(\tau_0,\tau_1,\tau_2)} \rho_{\text{covalent}}(\mathbf{r})\,d\mathbf{r}, \quad \tau_0 \geq \rho_a. \tag{11d}$$

Since the spatial integration (11*d*) discards the information about molecular conformation, $N_{\text{covalent}}(\tau_0, \tau_1, \tau_2)$ depends on only the covalent structure, *i.e.* covalent bonds and angles between atoms in the molecule, represented by $\rho_{\text{covalent}}(\mathbf{r})$ and is independent of molecular conformation. Therefore, from (11*b*), **EGL** is determined by chemical composition of a molecule and is independent of molecular conformation for $\tau_0 \geq \rho_a$.

In contrast, $N_{\text{nonbond}}(\tau_0, \tau_1, \tau_2)$ is determined by molecular packing as shown in (11*c*). The shape of $\rho_{\text{nonbond}}(\mathbf{r})$ therein depends on the atom types and packing configurations of neighboring, but non-bonded atoms. However, $N_{\text{nonbond}}(\tau_0, \tau_1, \tau_2)$ won't vary greatly for different proteins because the atoms of O, C and N which are the principle components of a protein have similar atom types, and because globular proteins have in common a close-packed configuration (Chothia, 1975; Richards, 1977; Ponder & Richards, 1987; Harpaz, Gerstein & Chothia, 1994). Moreover, the feature of local packing is smeared by the spatial average in (11*c*). Therefore, from (11*a*), **EGL** is also approximately irrelevant to molecular conformation for $\tau_0 < \rho_a$.

We have shown that the histogram $\mathbf{EGL}(\tau_0, \tau_1, \tau_2)$ is determined largely by the chemical composition of a molecule and should be relatively independent of the molecular conformation. In the following, we will verify these conclusions using simulated models.

3. Tests and discussion

In the tests, the electron density $\rho(\mathbf{r})$ was calculated using the *FFT* program of *CCP*4 (Collaborative Computational Project, Number 4, 1994). The magnitude of the gradient $|\nabla\rho(\mathbf{r})|$ and the Laplacian $\nabla^2\rho(\mathbf{r})$ are also calculated using the *FFT* program with the Fourier coefficients modified accordingly,

$$|\nabla\rho(\mathbf{r})| = \left\{[\partial\rho(\mathbf{r})/\partial x]^2 + [\partial\rho(\mathbf{r})/\partial y]^2 + [\partial\rho(\mathbf{r})/\partial z]^2\right\}^{1/2}, \tag{12a}$$

$$[\partial\rho(\mathbf{r})/\partial x] = -2\pi\sum_{hkl} i(a_1 h)\mathbf{F}_{hkl} \times \exp[-2\pi i(hx_1 + ky_1 + lz_1)], \tag{12b}$$

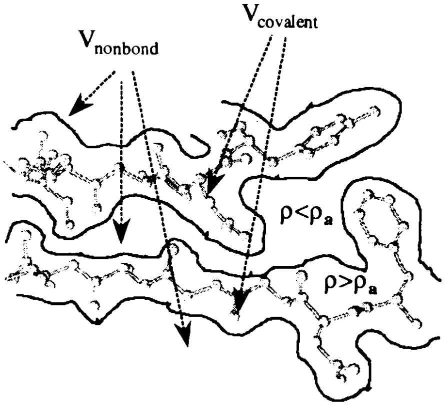


Fig. 1. Schematic drawing of V_{nonbond} and V_{covalent}. The molecular volume is divided into two regions, V_{nonbond} ($\rho < \rho_a$) and V_{covalent} ($\rho \geq \rho_a$). V_{covalent} is the volume occupied by main-chain and side groups of the molecule and V_{nonbond} is contributed from nearby residues that interact with each other either by van der Waals contacts or hydrogen bonds. The molecular model is taken from the crystal structure of cytidine deaminase (Betts, Xiang, Short, Wolfenden & Carter, 1994).

$$[\partial\rho(\mathbf{r})/\partial y] = -2\pi \sum_{hkl} i(a_2h + b_2k)\mathbf{F}_{hkl} \times \exp[-2\pi i(hx_1 + ky_1 + lz_1)], \quad (12c)$$

$$[\partial\rho(\mathbf{r})/\partial z] = -2\pi \sum_{hkl} i(a_3h + b_3k + c_3l)\mathbf{F}_{hkl} \times \exp[-2\pi i(hx_1 + ky_1 + lz_1)], \quad (12d)$$

$$\nabla^2\rho(\mathbf{r}) = -4\pi^2 \sum_{hkl} D_{hkl}\mathbf{F}_{hkl} \times \exp[-2\pi i(hx_1 + ky_1 + lz_1)], \quad (13a)$$

$$D_{hkl} = (a_1^2 + a_2^2 + a_3^2)h^2 + (b_2^2 + b_3^2)k^2 + c_3^2l^2 + 2(a_2b_2 + a_3b_3)hk + 2a_3c_3hl + 2b_3c_3kl, \quad (13b)$$

where the elements of orthogonalization matrix are $a_1 = a^*\sin(\beta^*)\sin(\gamma)$, $a_2 = -a^*\sin(\beta^*)\cos(\gamma)$, $a_3 = a^*\cos(\beta^*)$, $b_2 = b^*\sin(\alpha^*)$, $b_3 = b^*\cos(\alpha^*)$ and $c_3 = c^*$. The a^*, b^*, c^*, α^*, β^* and γ^* are reciprocal cell parameters and x_1, y_1 and z_1 crystallographic coordinates. The orthogonal axes **x**, **y** and **z** were chosen such that **x** along the crystallographic axis **a** and **z** along **c***.

3.1. *Independence of molecular conformation*

To verify the insensitivity of histograms to molecular conformation, we built three different secondary structures artificially from the same 16-residue peptide. For simplicity we only test the one-dimensional histograms **E**, **G** and **L**.

An α-helix of 16 residues taken from the cytidine deaminase crystal structure (Betts, Xiang, Short, Wolfenden & Carter, 1994) was used as one of the test models. The other models, a β-strand and a loop were built from the same residues to generate different conformations. The geometry of each model was refined with *REFI* in *FRODO* (Jones, 1985). An artificial space group $P3_121$ and unit cell of $a = 60.0$, $b = 60.0$, $c = 75.0$ Å, $\alpha = 90.0$, $\beta = 90.0$ and $\gamma = 120.0°$ were used throughout the calculation. The molecular envelopes *i.e.* the V_T's within which the histograms were evaluated were calculated from the corresponding structure models. To evaluate the influences of the stereochemical information on the histograms, random atom structure models were also generated by assigning random positions to the atoms of the 16 residues such that the atomic centers remained inside the corresponding molecular envelopes of the different secondary structures. Their molecular envelopes were then determined from the corresponding random atom models.

All the calculations were performed at a resolution of 2.0 Å which is often attainable for the X-ray diffraction from protein crystals, and which includes the most crucial reflections for the initial phase determination. The histograms depend on the resolution of the electron-density distribution and are more sensitive to structural features at the higher resolution. Thus, if the histograms prove identical at this resolution, the same will be true at lower resolution.

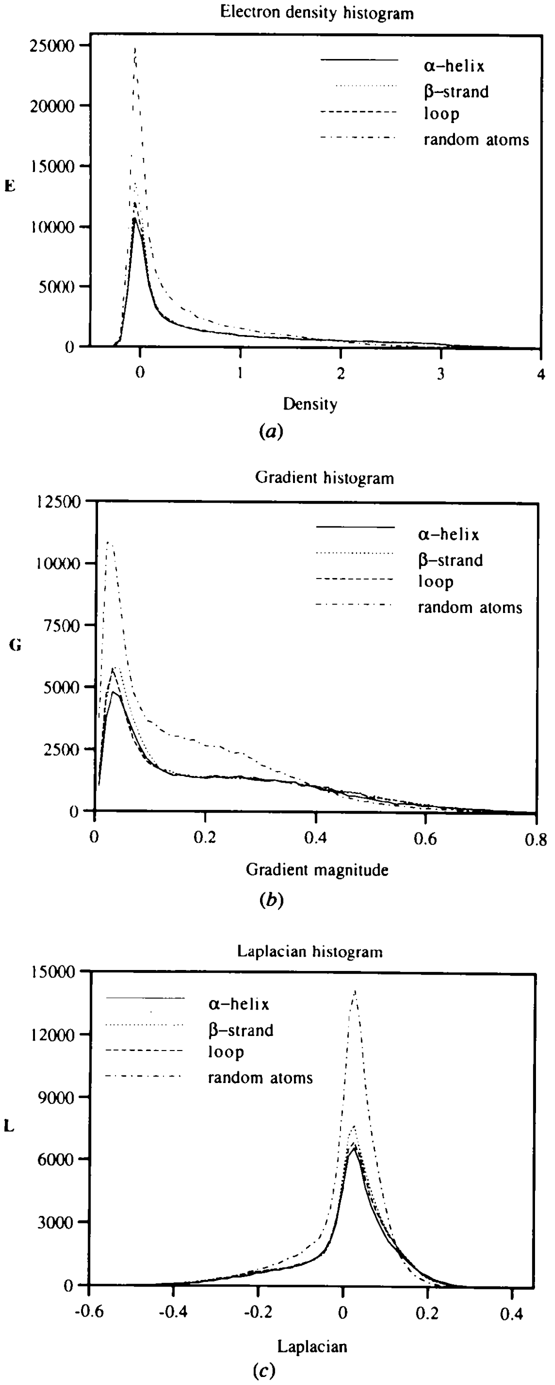


Fig. 2. Comparisons between one-dimensional histograms calculated from a random atom structure and the secondary structures of an α-helix, a β-strand and a loop at 2.0 Å resolution. The secondary structures consist of the same 16 residues with standard stereochemical geometry. The random atom structure is generated using the same atoms in the 16 residues.

The electron-density histograms (Fig. 2*a*) of the three secondary-structure models are almost the same in the high-density region and are different in the low-density region. The low-density region is contributed by electron densities from V_{nonbond} and, therefore, reflects the different molecular conformations. The high-density region is contributed by electron densities from the volume V_{covalent} and thus, represents the stereochemical properties of covalent bonds and angles.

Similarly, the gradient histograms (Fig. 2*b*) are almost identical for the different secondary structures throughout the entire range of the gradient magnitude, $|\nabla\rho(\mathbf{r})|$, except where its value approaches zero. The largest gradient values derive from the region close to inflection points in the density map and these all lie at some small distance from atomic centers and from the axes of chemical bonds between them within V_{covalent}. Therefore, these high values are derived from the atom types and covalent structure of the molecule. The low gradient values are contributed mainly by electron density in the volume V_{nonbond} and thus represent information about packing interactions.

The Laplacian histograms for the three secondary structures are shown in Fig. 2(*c*). They too are much the same except for a small difference observed in the strong peak located close to 0, a region also contributed mostly by electron densities in V_{nonbond}. This peak represents contributions from the molecular folding. The long tail in the negative region results mainly from the electron density in V_{covalent} near or between atomic centers, and thus reflects stereochemical information about covalent bonds and angles.

In contrast, the histograms of the random atom structures differ significantly everywhere from those of the corresponding secondary structures. The histograms from the random atom model generated inside the helix envelope are also shown in Fig. 2. The large discrepancies arise from the fact that the secondary structures have correct bond lengths and angles whereas the random atom structure does not. It should be noted that the most significant deviations occur in the gradient histogram which differs considerably from the histograms of the secondary structures not only at small gradient values but also at the region of medium values. This is because the gradient histogram encodes much more stereochemical information than either the electron density or the Laplacian histograms, as discussed below.

In summary, histograms calculated from different conformations with different secondary structures are almost the same in regions contributed by electron densities in V_{covalent}, and differ slightly in regions to which electron densities in V_{nonbond} make major contributions. In contrast to the relative independence from molecular conformation, the proper stereochemical bonding between atoms has significant effects on the histograms. These observations indicate that the histograms defined here are mainly determined by stereochemical properties of the covalent structure of polypeptides, and relatively independent of molecular conformations.

3.2. *Sensitivity to phase errors*

The usefulness of the histograms in density modification and *ab initio* phasing applications depends on their sensitivity to phase errors. The electron-density histogram has shown some phase-discriminative capability, as

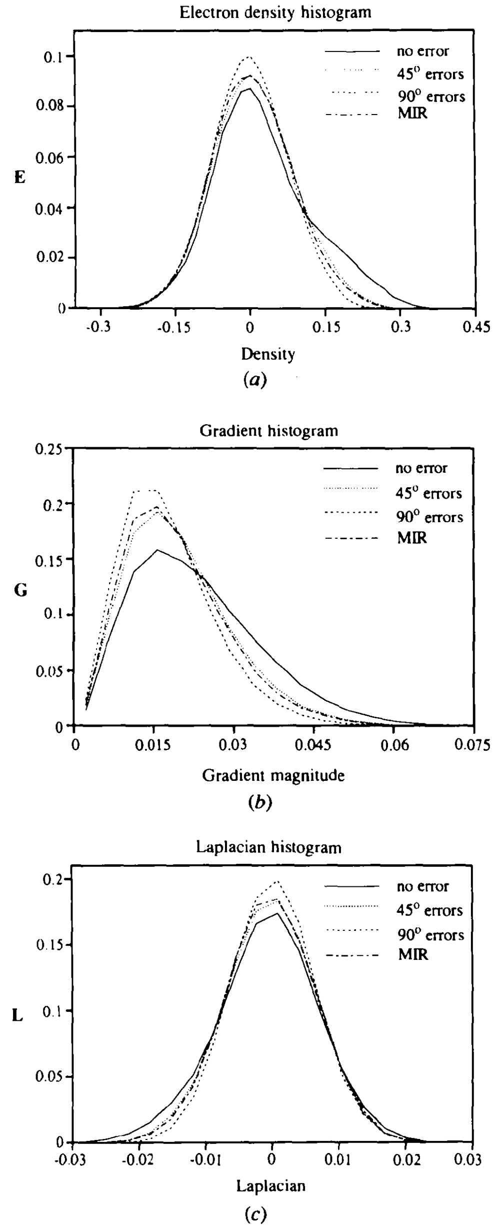
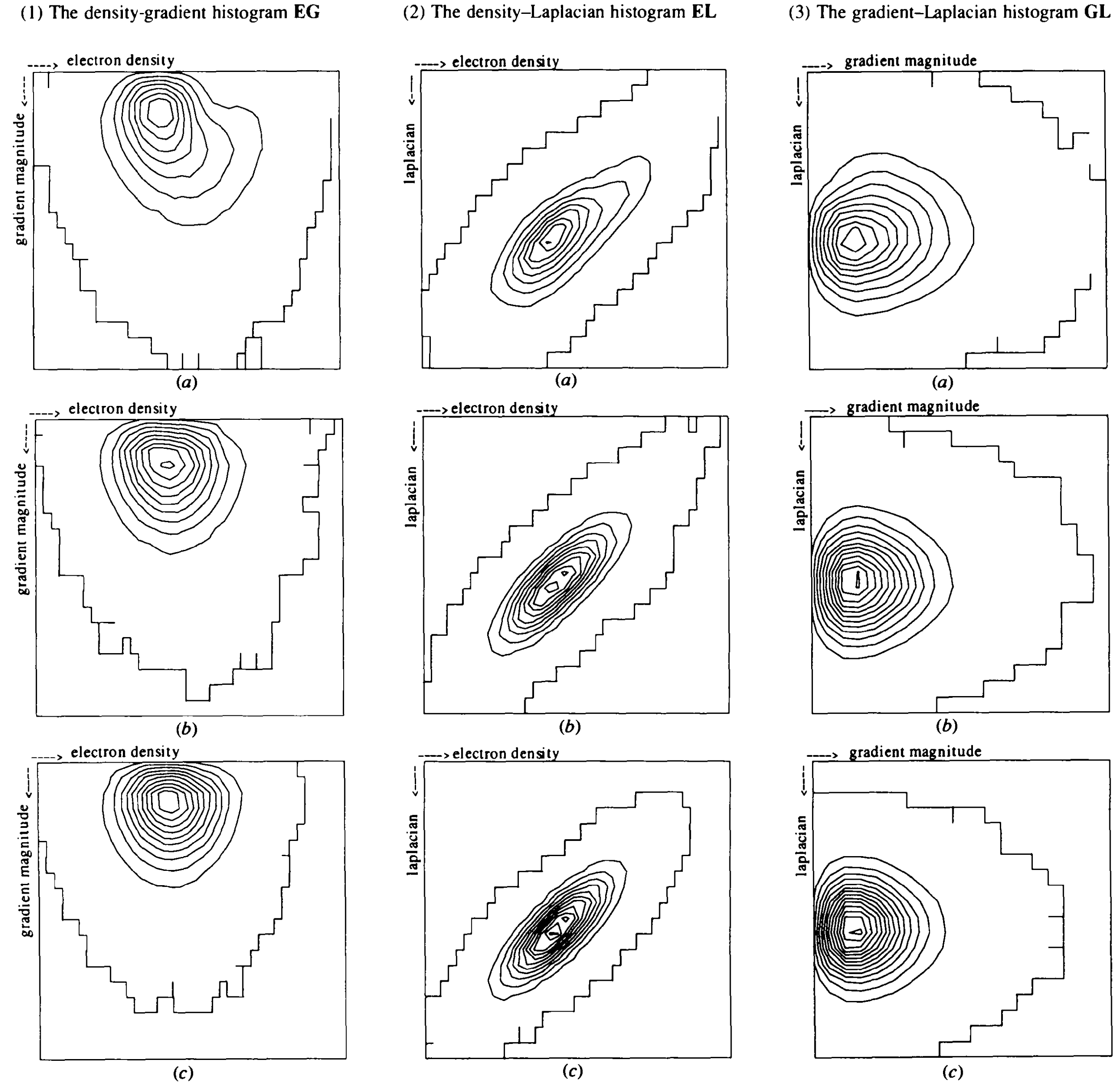


Fig. 3. One-dimensional histograms calculated for cyclophilin A at 3.5 Å resolution using MIR phases and phases with 0, 45 and 90° random phase errors.

indicated by its successful application in the phase extension and refinement for macromolecules (Zhang & Main, 1990*b*). The new histograms we proposed here encode more stereochemical information than the electron-density histogram alone and should be more sensitive to phase errors.

The phase sensitivity of the new histograms was tested using 3.5 Å X-ray diffraction data of cyclophilin A (Ke, Zydowsky, Liu & Walsh, 1991). Error-free histograms were calculated from the experimental amplitudes $|F_o|$ and phases φ_c obtained from the cyclophilin A structure model. Simulated-error histograms were generated using the same amplitudes $|F_o|$ and the phases $\varphi_c + \varphi_r$ with the random errors φ_r introduced. The real-error histograms were also calculated using $|F_o|$ and MIR phases. All the histograms were evaluated within the same molecular envelope determined from the refined model.

To give a quantitative measure of the sensitivity of the histograms to phase errors, we define the R factor of histograms as, $R_h = \sum |p - p_m| / \sum p_m$, where p is the


(1) The density-gradient histogram **EG**

(2) The density–Laplacian histogram **EL**

(3) The gradient–Laplacian histogram **GL**

Fig. 4. Two-dimensional histograms calculated for cyclophilin A at 3.5 Å resolution using (*a*) error-free phases, (*b*) MIR phases and (*c*) random phases.

histogram of the electron-density distribution in question and p_m the error-free histogram. R_h measures differences between a histogram and the error-free histogram and, by implication, the phase differences between them if the amplitudes of structure factors are specified as in this test.

The one-dimensional histograms, **E**, **G** and **L** are shown in Fig. 3. The electron-density histogram **E** (Fig. 3*a*) with 90° random phase errors is nearly symmetric. As phase errors decrease, the peak gets lower and broader with a tail extending to the region of high electron density. Similarly, in the gradient histogram **G** (Fig. 3*b*) and the Laplacian histogram **L** (Fig. 3*c*), the peaks become lower and broader with phase errors reduced and concomitantly the tails extend to the region where the densities in $V_{\rm covalent}$ makes major contributions. Among the three types of histograms, the gradient histograms **G** show the largest deviation between the error-free and the completely random phases, which is also indicated by the R_h in Fig. 5. The histograms **E**, **G** and **L** calculated from MIR phases are all located between the corresponding histograms with 45° and 90° phase errors. The average phase errors of the MIR phases would be estimated, based on the histograms, to be a little more than 45°.

The two-dimensional histograms, **EG**, **EL** and **GL** are shown in Fig. 4. A common feature is that the peak becomes broader, less condensed and more asymmetric as the phase errors decrease. This is correlated with the changes in one-dimensional histograms mentioned above.

The variation induced in R_h by phase errors indicates the phase sensitivity of the corresponding histogram. The R_h's calculated for the histograms are compared as a function of average phase error in Fig. 5. They all decrease monotonically as phase errors are reduced from 90 to 0°. In the case of one-dimensional histograms, the variations are 0.16, 0.36 and 0.17, respectively, for the histograms, **E**, **G** and **L** when phase errors increase from 0 to 90°. It is interesting that the gradient histogram **G** once again shows significantly more phase sensitivity compared to the density histogram **E** and the Laplacian histogram **L** as suggested by the comparison between the random atom structure and the secondary structures in §3.1 above.

The two-dimensional histograms have increased sensitivity to phase errors, and, correspondingly the variations for those involving the gradient, the **EG** (0.43) and **GL** (0.42), are greater than that for the **EL** histogram (0.29). The density-gradient histogram **EG** gives a higher variation (0.43) than its projections **E** (0.16) and **G** (0.36), and thus is more sensitive to phase errors. Similarly, the density–Laplacian histogram **EL** and the gradient–Laplacian histogram **GL** have higher phase sensitivity than their corresponding projections, the one-dimensional histogram **E**, **G** and **L** alone. The three-dimensional histogram **EGL** shows even larger phase-error dependent variation, 0.49 compared to those of the two-dimensional projections **EG** (0.43), **EL** (0.29) and **GL** (0.42). This indicates that the components of the histograms, the density $\rho(\mathbf{r})$, the magnitude of gradient $|\nabla\rho(\mathbf{r})|$ and Laplacian $\nabla^2\rho(\mathbf{r})$ encode somewhat independent stereochemical information. Therefore, the phase sensitivity of the multi-dimensional histogram can be enhanced by including additional components since these components are at least partially independent.

An important feature in Fig. 5 is that the phase-error dependencies for histograms containing the gradient magnitude are steeper than for those which do not. The observation that the gradient histogram is more sensitive both to stereochemical constraints and to phase errors than either the density or the Laplacian histograms, suggests that the gradient makes the most significant contributions to the phase sensitivity of the histograms containing it. Since the gradient measures differences between neighboring density values, it is more sensitive to detailed geometrical features of the electron-density distribution than is the density itself or the Laplacian. Therefore, the enhanced phase sensitivity of the histograms containing the gradient magnitudes arises probably because the gradient captures more stereochemical information owing to the higher molecular shape sensitivity compared to either the density or the Laplacian.

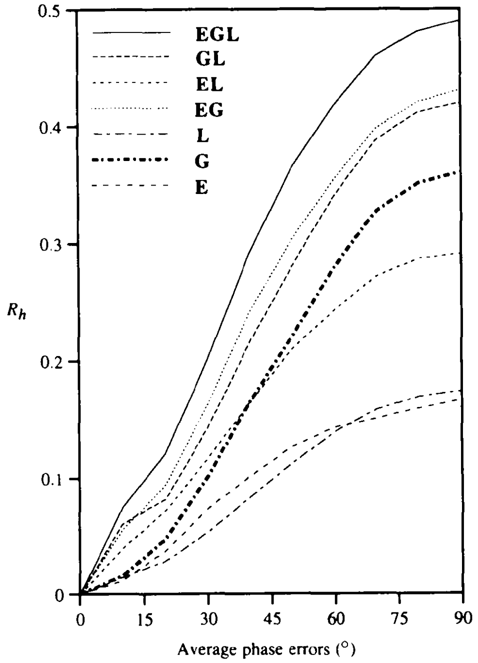


Fig. 5. Phase sensitivity of histograms: R_h plotted as a function of average phase errors, $R_h = \sum|p - p_m|/\sum p_m$, p is the histogram of an electron-density distribution in question and p_m the corresponding error-free histogram. All the histograms were calculated at 3.5 Å resolution.

4. Concluding remarks

The multi-dimensional histogram proposed here represents stereochemical properties encoded in the shape of the macromolecular electron-density distribution. It represents much more information than the electron-density histogram previously developed (Lunin, 1988; Zhang & Main, 1990*b*) since it reflects not only the frequency of a particular electron-density value but also its relationship with the densities of its neighbors. It is mainly dependent on chemical composition of a molecule and minimally dependent on molecular folding and packing.

The multi-dimensional histogram can be divided, as shown in the tests, into regions that represent stereochemical information about covalent bonds, bond angles and atom types, and regions that depend on molecular folding and packing. Thus, the former regions of the multi-dimensional histogram can be calculated accurately from the composition of a molecule and can serve as reliable targets in histogram matching. The latter regions can be estimated from known structures with little loss in precision since no dramatic changes would be expected between different macromolecular structures as discussed above. Therefore, they also can be used as targets but they should be down-weighted.

The multi-dimensional histograms are significantly more sensitive to phase errors than is the conventional electron-density histogram alone. Phase sensitivity can be enhanced by including higher order derivatives into the multi-dimensional histogram. The gradient histogram has much greater sensitivity to phase errors than does the electron-density histogram; thus, it should serve as a basis for much better figure of merit for selecting correct phase sets than the electron-density histogram in multi-solution phase determinations (Lunin *et al.*, 1990).

Like the electron-density histogram, the multi-dimensional histogram can be easily realized in computer algorithms. The calculations of the gradients of an electron-density map require three Fourier transformations, as indicated in (12). the efficiency of the calculation can be enhanced by using the fast Fourier transformation (FFT) algorithm (Gentleman & Sande, 1966; Ten Eyck, 1973; Brünger, 1989; An, Lu, Prince & Tolimieri, 1992; Bricogne, 1993). The Laplacian calculations are rather simple, requiring just one FFT [(13)]. It should be noted that it is possible to use higher dimensional histograms but with more computing costs. Since the even-order derivatives can be calculated by only one FFT, they could be used to construct the multi-dimensional histogram with increasing computing efficiencies.

Further studies on the calculation of the multi-dimensional histogram from macromolecular chemical compositions and similar structures and its application to improve macromolecular electron-density maps are under investigation.

One of the authors, Shibin Xiang, thanks Dr Lee Kuyper for reviewing the paper and kind suggestions, and Dr Hengming Ke for providing the cycophilin A data.

References

An, M., Lu, C., Prince, E. & Tolimieri, R. (1992). *Acta Cryst.* A**48**, 415–418.

Betts, L., Xiang, S., Short, S. A., Wolfenden, R. & Carter, C. W. Jr (1994). *J. Mol. Biol.* **235**, 635–656.

Borden, S. R. (1983). *A Course in Advanced Calculus.* New York: North Holland.

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

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

Bricogne, G. (1993). *Fourier Transforms in Crystallography: Theory, Algorithms, Applications*, in *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, pp. 23–106. Dordrecht: Kluwer Academic Publishers.

Brünger, A. (1989). *Acta Cryst.* A**45**, 42–50.

Brünger, A. (1992). *X-PLOR Version* 3.0 *Manual.* Yale University, New Haven, CT, USA.

Cannillo, E., Oberti, R. & Ungaretti, L. (1983). *Acta Cryst.* A**39**, 68–74.

Chothia, C. (1975). *Nature (London)*, **254**, 304–308.

Cochran, W. (1952). *Acta Cryst.* **5**, 65–67.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Collins, D. M. (1978). *Acta Cryst.* A**34**, 533–541.

Collins, D. M., Brice, M. D., La Cour, T. F. M. & Legg, M. J. (1976). *Crystallographic Computing Techniques*, pp. 330–335. Copenhagen: IUCr/Munksgaard.

Gentleman, W. M. & Sande, G. (1966). Proceedings of the Fall Joint Computer Conference 1966, IEEE Computer Society, pp. 563–578.

Harker, D. & Kasper, J. S. (1948). *Acta Cryst.* **1**, 70–75.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Structure*, **2**, 641–649.

Harrison, R. W. (1988). *J. Appl. Cryst.* **21**, 949–952.

Hauptman, H. (1986). *Science*, **233**, 178–183.

Hendrickson, W. A. (1985). *Methods Enzymol.* **115**, 252–270.

Jones, A. (1985). *Methods Enzymol.* **115**, 157–171.

Karle, J. (1986). *Science*, **232**, 837–843.

Karle, J. & Hauptman, H. (1950). *Acta Cryst.* **3**, 181–187.

Ke, H. M., Zydowsky, L. D., Liu, J. & Walsh, C. T. (1991). *Proc. Natl Acad. Sci. USA*, **88**, 9483–9487.

Lunin, V. Y. (1988). *Acta Cryst.* A**44**, 144–150.

Lunin, V. Y. (1993). *Acta Cryst.* D**49**, 90–99.

Lunin, V. Y., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.* A**46**, 540–544.

Main, P. (1990). *Acta Cryst.* A**46**, 507–509.

Ponder, J. W. & Richards, F. M. (1987). *Cold Spring Harbor Symp. Quant. Biol.* **LII**, 421–428.

Richards, F. M. (1977). *Ann. Rev. Biophys. Bioeng.* **6**, 151–176.

Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.

Sayre, D. (1974). *Acta Cryst.* A**30**, 180–184.

Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.

Tronrud, D., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.

Woolfson, M. M. (1987). *Acta Cryst.* A**43**, 593–612.

Zhang, K. Y. J. (1993). *Acta Cryst.* D**49**, 213–222.

Zhang, K. Y. J. & Main, P. (1990*a*). *Acta Cryst.* A**46**, 377–381.

Zhang, K. Y. J. & Main, P. (1990*b*). *Acta Cryst.* A**46**, 41–46.