Acta Cryst. (1991). A47, 45-52

Frequency-Restrained Structure-Factor Refinement. I. Histogram Simulation

By V. Yu. Lunin and T. P. Skovoroda

Research Computing Center, USSR Academy of Sciences, Pushchino, Moscow region 142292, *USSR*

(*Received* 19 *May* 1989; *accepted* 4 *September* 1990)

Abstract

An analysis of the frequencies of different values encountered in protein electron-density syntheses reveals characteristic shapes for their distributions (histograms). This property can be used to refine ill-defined phases (and, perhaps, some of the moduli) of structure factors, and thus to obtain more-interpretable electron-density maps. A simple empirical model is designed which can predict the histogram for a protein with an undetermined structure provided its unit-cell volume and charge are known. The parameters of the histogram model are derived from a set of proteins with known spatial structures. The application of the simulated histogram is illustrated by an improved electron-density map for the 'dry' form of the protein γ-crystallin IIIb.

1. Introduction

1.1. *Imaging electron-density distribution in a protein*

Normally, the X-ray determination of the spatial structure of a biological macromolecule starts with searching for a function $\rho(\mathbf{r})$ which could characterize the electron-density distribution in the whole crystal. In practice, this search always reduces to an approximation of this function, $\rho_d(\mathbf{r})$, at a finite resolution d:

$$\rho_d(\mathbf{r}) = |V|^{-1} \sum_{|\mathbf{s}|\leq 1/d} F(\mathbf{s}) \exp[i\varphi(\mathbf{s})] \times \exp[-2\pi(\mathbf{s}, \mathbf{r})]. \quad (1)$$

We call $\rho_d(\mathbf{r})$ the 'image' of $\rho(\mathbf{r})$ at a resolution d. Again, this image may not always be accurate enough to be interpreted properly, since some of the phases $\varphi(\mathbf{s})$, and even some of the moduli $F(\mathbf{s})$, may either be approximate or not known. Then, to remedy the situation and improve the image, one should have recourse to other types of information on the object studied, which could provide more-accurate values of structure factors to calculate the synthesis (1).

1.2. *Restrictions on the range of* $\rho(\mathbf{r})$ *values*

Various restrictions on the range of $\rho(\mathbf{r})$ are widely used examples of such additional information. The function may be assumed to be non-negative [$\rho(\mathbf{r}) \geq 0$], bounded on both sides [$\rho_{\min} \leq \rho(\mathbf{r}) \leq \rho_{\max}$ with $\rho_{\min}$ and $\rho_{\max}$ preset], having a finite set of values [$\rho(\mathbf{r}) = \{0 \text{ or } 1\}$] *etc.*; the examples are many. Efforts were made to modify structure-factor phases in (1) in order that $\rho_d(\mathbf{r})$ possessed these properties.

To examine more thoroughly the range of values the functions $\rho(\mathbf{r})$ and $\rho_d(\mathbf{r})$ can take, one should not only find out which they are, but also how many times each of them occurs. In this way, one arrives at the conclusion that the properties of the actual function $\rho(\mathbf{r})$ and of its image $\rho_d(\mathbf{r})$ generally are different.

Assume for simplicity that a function $f(\mathbf{r})$ [or some other one, $\rho(\mathbf{r})$ or $\rho_d(\mathbf{r})$ for instance] is calculated at N_{tot} grid points. Break an interval $(\rho_{\min}, \rho_{\max})$ of the real axis into K portions (bins). Let n_k be the number of grid points where $f(\mathbf{r})$ values fall in the kth bin ($k = 1, \ldots, K$). Define the normalized frequencies that these values occur in the bins to be

$$\nu_k = (1/\Delta_k)(n_k/N_{\text{tot}}), \quad (2)$$

where Δ_k is the length of the kth bin.

We call the set of normalized frequencies $\{\nu_k\}_{k=1}^{K}$ the histogram of $f(\mathbf{r})$.

An analysis of experimental data (Podjarny & Jonath, 1977; Lunin, 1986, 1988; Zhang & Main, 1990) has shown that histograms of the images $\rho_d(\mathbf{r})$ for protein electron densities have specific asymmetrical shapes (Fig. 1). They are noticeably distinct from the histograms of synthesis (1) with random phases or with a number of excluded reflections (Lunin, 1988). Note that the shape of the histogram of $\rho_d(\mathbf{r})$ depends not only on the properties of the corresponding function $\rho(\mathbf{r})$ but also on the resolution d.

1.3. *Improving the images by restraining the frequencies*

The histograms of images of electron-density distributions in proteins may be used to improve the quality of synthesis (1) (Lunin, 1986, 1988; Harrison, 1988; Zhang & Main, 1990). Assume that the histogram $\{\nu_k^0\}_{k=1}^{K}$ for the desired function $\rho_d(\mathbf{r})$ is known. Then we choose the values of the unknown structure-factor phases (and, perhaps, a small number of unknown moduli too) so as to minimize the discrepancy

$$\sum_{k=1}^{K} w_k(\nu_k^c - \nu_k^o)^2 \Rightarrow \min, \qquad (3)$$

or any other similar criterion. Here ν_k^c are the normalized frequencies for an image $\rho_d^c(\mathbf{r})$ with some 'trial' values of the unknown structure factors, and w_k are prescribed weights.

It was demonstrated with tests that the approach considerably improves the image $\rho_d(\mathbf{r})$.

The specificity of histograms (or of their characteristics) was used as a criterion for a proper choice of a phase set by Luzzati, Mariani & Delacroix (1988), Mariani, Luzzati & Delacroix (1988) and Lunin, Urzhumtsev & Skovoroda (1990). A close, though reciprocal-space, approach was suggested by Hašek and his colleagues to select the best phases from some allowable ones (Hašek, 1984; Hašek, Schenk, Riers & Schgen, 1985; Hašek & Schenk, 1988; Kříž, 1989). The criterion was the degree of agreement between the 'empirical' and the 'theoretical' distributions of semiinvariants.

A different approach to using specific features of histograms was proposed (Harrison, 1988; Zhang & Main, 1990). It consisted in modifying the synthesis (1) to one with a 'good' histogram, and using the phases calculated from the modified synthesis for further iterative phase improvement. The relationship between the methods of 'phase refinement' (3), 'histogram matching' (Zhang & Main, 1990) and 'density modification' will be examined elsewhere (Lunin & Vernoslova, 1990).

1.4. *Prediction of histograms*

The aim of this paper is to show a way of simulating the 'standard' histogram $\{\nu_k^0\}_{k=1}^K$ for a protein with an unknown structure. First, we propose in § 2.1 a formula to describe shapes of histograms corresponding to proteins. This formula contains a number of parameters whose values must be determined. We define these parameters so that 'theoretical' curves calculated with the formula are in good agreement with the true histograms for a lot of proteins with known three-dimensional structure. The formula with the determined parameters may be used then for the prediction of histograms for proteins with unknown spatial structure. The simple empirical model we propose here does not claim to reproduce all the fine points of a protein histogram. However, as shown by the example in § 3, the accuracy of simulated histograms is good enough for them to be used successfully in practice.

1.5. *Short mathematical description*

The normalized frequencies (2) depend, in the strict sense, on the grid where the values of $\rho_d(\mathbf{r})$ are calculated and on the bin lengths Δ_k. To be more accurate, we should examine a limit case when the number of grid points grows and the bin lengths tend to zero:

$$\nu(t) = \lim_{\Delta \to 0} (1/\Delta)\{\text{the volume of the part of the unit cell where } t - \Delta/2 \le \rho_d(\mathbf{r}) \le t + \Delta/2\} \times \{\text{the unit-cell volume}\}^{-1}. \qquad (4)$$

Here the function $\nu(t)$ depends only on the image $\rho_d(\mathbf{r})$, not on the choice of the grid or the bin length. The values of ν_k in (2) approximate those, $\nu(t_k)$, that the function $\nu(t)$ takes in the middle of the bin, t_k. By a histogram we will mean not only the set of normalized frequencies $\{\nu_k\}_{k=1}^K$, but also the function $\nu(t)$ whose value these frequencies approximate.

The function $\nu(t)$ is such that

$$\int_{-\infty}^{\infty} \nu(t)\,\mathrm{d}t = 1, \qquad V\int_{-\infty}^{\infty} t\nu(t)\,\mathrm{d}t = F_{000},$$

where F_{000} is the full charge of the unit cell, and V is its volume.

The discrete analogs of these properties are

$$\sum_{k=1}^{K} \nu_k\Delta_k = 1, \qquad \sum_{k=1}^{K} t_k\nu_k\Delta_k = F_{000}/V, \qquad (5)$$

where t_k are the bin middles.

2. Simulating the histogram

The first question which immediately suggests itself when one tries to use the criterion (3) is what is the standard histogram $\{\nu_k^0\}_{k=1}^K$ for the unknown image?

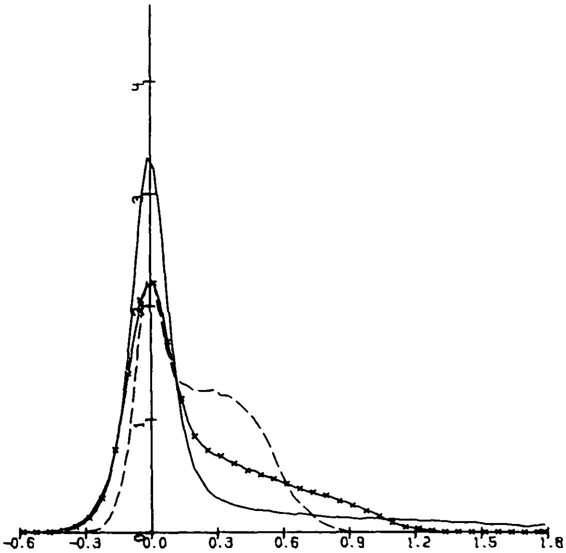


Fig. 1. Histograms of electron-density-distribution images in a cytochrome crystal at various resolutions (—— 2 Å; —×— 4 Å; – – 6 Å).

Here we describe the empirical procedure for simulating histograms for proteins with undetermined structures from histograms corresponding to proteins with known three-dimensional structures.

The shape of the histogram for $\rho_d(\mathbf{r})$ depends on image resolution d (Fig. 1). In this paper we restrict ourselves to images at a resolution of 4 Å. An analogous procedure can be applied at any other medium or high resolution.

The shape of a histogram also depends on the atomic temperature factors. Fig. 2 shows histograms corresponding to different values of the temperature factor. In all other calculations we put the values of the temperature parameters for all atoms equal to 10 Å^2.

This paper is devoted to a computer analysis of known protein structures. Therefore, wherever images of known proteins are encountered we deal with those images (1) whose structure factors are calculated from atomic coordinates. These were taken from the Protein Data Bank.

2.1. *Empirical histogram model*

Fig. 3 shows histograms corresponding to different proteins. It can be seen that the frequencies (2) vary markedly from protein to protein. But the situation changes when we move to the 'normalized volumes', so that

$$v_k = \nu_k V / F_{000} \qquad (6)$$

[v_k is the volume of the part of the unit cell where the values of the function $\rho_d(\mathbf{r})$ lie in the kth bin corresponding to one electron of charge].

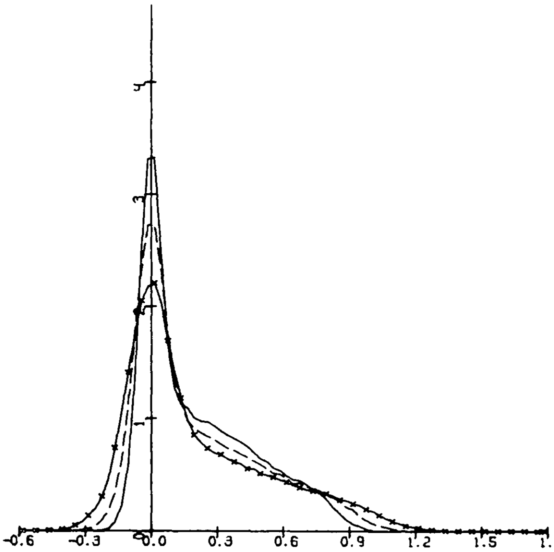


Fig. 2. The histograms of 4 Å electron-density-distribution images in a cytochrome crystal at various atomic temperature factors (—— $B = 50$ Å^2; — — $B = 30$ Å^2; —×— $B = 10$ Å^2).

Analogously, we introduce the function

$$v(t) = \nu(t) V / F_{000}.$$

Fig. 4 shows the curves of Fig. 3 modified by (6). Here we see that the plots are much the same in the left-hand and right-hand parts, showing considerable differences in the region of the central peak only. This may be explained by the proteins having more- or less-compact molecular packing and, therefore,

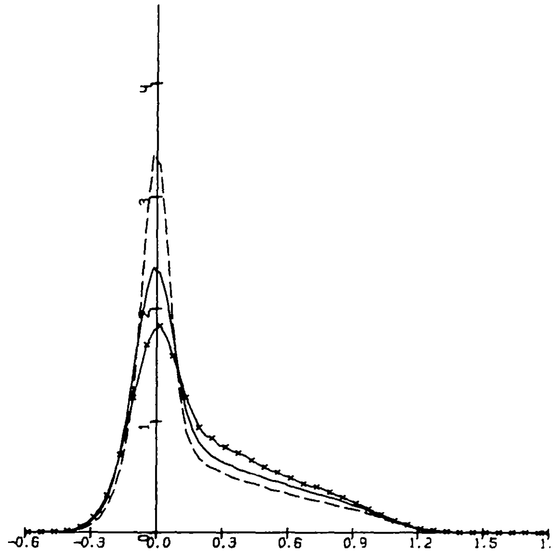


Fig. 3. The histograms of 4 Å electron-density-distribution images in various protein crystals (—— staphylococcal nuclease; — — chymotrypsinogen; —×— carbonic anhydrase).

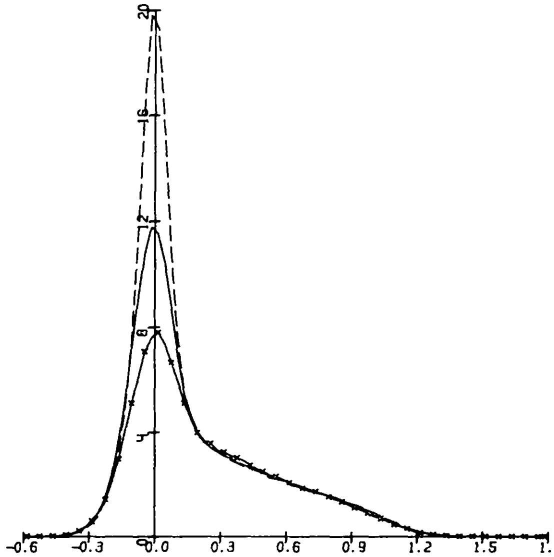


Fig. 4. The modified histograms (see § 2.1) of 4 Å electron-density-distribution images in various protein crystals (—— staphylococcal nuclease; — — chymotrypsinogen; —×— carbonic anhydrase).

different volumes of unoccupied lower electron-density regions.

The simplest empirical model which we suggest to describe the modified histograms $\{v_k\}_{k=1}^K$ has the form

$$v_k = v_k^0 + \left(V/F_{000} - \sum_{j=1}^{K} v_j^0 \Delta_j \right) q_k^0, \quad k = 1, \ldots, K. \tag{7}$$

Here, $\{v_k^0\}_{k=1}^K$ is a universal distribution of the normalized volumes, common for all proteins and 'responsible' for the distribution of $\rho_d(\mathbf{r})$ values inside the molecule region; and $\{q_k^0\}_{k=1}^K$ is a universal (common for all proteins) distribution 'responsible' for the distribution of $\rho_d(\mathbf{r})$ values in the protein-free volume of the unit cell. This volume per single electron charge is

$$V/F_{000} - \sum_{j=1}^{K} v_j^0 \Delta_j.$$

Model (7) is a discrete analog of the 'continuous' version

$$v(t) = v^0(t) + [V/F_{000} - \int_{-\infty}^{\infty} v^0(x)\,\mathrm{d}x] q^0(t).$$

2.2. *Model parameters v_k^0 and q_k^0*

The normalized frequencies ν_k obey (5). This means that if we want the model (7) to work for all F_{000} and V, we should require from $\{v_k^0\}_{k=1}^K$ and $\{q_k^0\}_{k=1}^K$ that for all F_{000} and V

$$\sum_{k=1}^{K} [v_k^0 + (V/F_{000} - \sum_{j=1}^{K} v_j \Delta_j) q_k^0] \Delta_k = V/F_{000}$$

and

$$\sum_{k=1}^{K} t_k [v_k^0 + (V/F_{000} - \sum_{j=1}^{K} v_j \Delta_j) q_k^0] \Delta_k = 1.$$

If we introduce the notation

$$S_v = \sum_{k=1}^{K} v_k^0 \Delta_k, \qquad R_v = \sum_{k=1}^{K} t_k v_k^0 \Delta_k,$$

$$S_q = \sum_{k=1}^{K} q_k^0 \Delta_k, \qquad R_q = \sum_{k=1}^{K} t_k q_k^0 \Delta_k,$$

we can rewrite these requirements as

$$\begin{aligned} F_{000} S_v (1 - S_q) + V(S_q - 1) &= 0 \\ F_{000}(R_v - S_v R_q - 1) + V R_q &= 0. \end{aligned} \tag{8}$$

These are valid for all F_{000} and V only when all their coefficients are zero, which means that the distributions $\{v_k^0\}_{k=1}^K$ and $\{q_k^0\}_{k=1}^K$ should be such that

$$\begin{gathered} S_q = \sum_{k=1}^{K} q_k^0 \Delta_k = 1, \qquad R_v = \sum_{k=1}^{K} t_k v_k^0 \Delta_k = 1, \\ R_q = \sum_{k=1}^{K} t_k q_k^0 \Delta_k = 0. \end{gathered} \tag{9}$$

Note that equations (8) do not fix the value S_v. Moreover, if we rewrite (7) as

$$v_k = (v_k^0 - S_v q_k^0) + (V/F_{000}) q_k^0, \quad k = 1, \ldots, K, \tag{10}$$

we can show that if (9) and (10) hold for some $\{v_k^0\}_{k=1}^K$ and $\{q_k^0\}_{k=1}^K$, they will also hold for v_k^0 replaced by $v_k^0 + \lambda q_k^0$ with $k = 1, \ldots, K$ and any λ. This means that $\{v_k^0\}_{k=1}^K$ in (7) may be ambiguous: when we change v_k^0 to $v_k^0 + \lambda q_k^0$ in (7), the values of v_k calculated from this formula remain the same. To get rid of the ambiguity, we can somehow fix a value of S_v, say, by putting

$$S_v = \sum_{k=1}^{K} v_k^0 \Delta_k = 0. \tag{11}$$

We stress that, unlike (9) which follows from (5), (11) is arbitrary, introduced to fix one of the possible parameter sets $\{v_k^0\}_{k=1}^K$. We equally could have claimed that S_v be equal to some other value.

2.3. *Determining histogram parameters*

We have found the parameters $\{v_k^0\}_{k=1}^K$ and $\{q_k^0\}_{k=1}^K$ in (7) from a set of $J(=15)$ proteins (Table 1) from the Protein Data Bank, which we call 'base' proteins. For each of them we used atomic coordinates to calculate structure factors, calculate the image $\rho_d(\mathbf{r})$ at a resolution of 4 Å and determine $\{v_k^{(j)}\}_{k=1}^K$ by (2)-(6). [Here the upper index (j) is the number of the protein in the base set, and k is the bin number.] The values of $\{v_k^0\}_{k=1}^K$ and $\{q_k^0\}_{k=1}^K$ were derived from the requirement that

$$v_{k,\mathrm{theor}}^{(j)} = v_k^0 + (V^{(j)}/F_{000}^{(j)} - \sum_{i=1}^{K} v_i^0 \Delta_i) q_k^0 \tag{12}$$

fit best the histograms of the base proteins:

$$Q = \sum_{j=1}^{J} \sum_{k=1}^{K} [N_{\mathrm{tot}}^{(j)} F_{000}^{(j)} \Delta_k / V^{(j)}][(v_{k,\mathrm{theor}}^{(j)} - v_k^{(j)})^2 / v_k^{(j)}] \Rightarrow \min \tag{13}$$

under additional conditions

$$\begin{gathered} \sum_{k=1}^{K} v_k^0 \Delta_k = 0, \qquad \sum_{k=1}^{K} t_k v_k^0 \Delta_k = 1, \\ \sum_{k=1}^{K} q_k^0 \Delta_k = 1, \qquad \sum_{k=1}^{K} t_k q_k^0 \Delta_k = 0. \end{gathered} \tag{14}$$

Here $F_{000}^{(j)}$, $V^{(j)}$ and $N_{\mathrm{tot}}^{(j)}$ are the charge, the volume and the number of grid points in the unit cell for the jth protein, respectively, and Δ_k is the length of the kth bin. The weight multiplier in the minimized criterion (13) represents passing over from values v_k to the numbers n_k of the grid points with the values in the kth bin. Hence, the minimized criterion (13) is merely

$$Q = \sum_{j=1}^{J} \sum_{k=1}^{K} (n_{k,\mathrm{theor}}^{(j)} - n_k^{(j)})^2 / n_k^{(j)},$$

Table 1. *Base protein set*

				Criteria of model and real histograms agreement	
Protein	Files*	NRes†	Reference	Q‡	Q_g§
Carbonic anhydrase B	2CAB	261	Kannan *et al.* (1975)	1649	1·15
Chymotrypsinogen A	1CHG	245	Freer, Kraut, Robertus, Wright & Xuong (1970)	1768	0·94
Cytochrome B5	2B5C	93	Mathews, Levine & Argos (1971)	3500	0·93
HIPIP	1HIP	85	Freer, Alden, Carter & Kraut (1975)	692	0·52
Bence-Jones protein	2PHE	114	Furey, Wang, Yoo & Sax (1983)	1417	1·14
Insulin	1INS	21+30	Dodson, Dodson, Hodgkin & Reynolds (1979)	4260	1·47
Lysozyme	1LZ1	130	Artymiuk & Blake (1981)	1303	0·54
Ovomucoid third domain	1OVO	4×56	Papamokos *et al.* (1982)	1788	0·71
Phospholipase	1BP2	123	Dijkstra, Kalk, Hol & Drenth (1981)	865	0·65
Plastocyanin	1PSY	99	Guss & Freeman (1983)	1221	0·89
Prealbumin	2PAB	2×127	Blake, Geisow, Oatley, Rerat & Rerat (1978)	1240	0·55
Proteinase A	2SGA	181	Sieleski *et al.* (1979)	2603	1·29
Ribonuclease A	1RN3	124	Borkakoti, Moss & Palmer (1982)	2140	0·68
Staphylococcal nuclease	2SNS	149	Cotton Hazen & Legg (1979)	7180	1·55
Myoglobin	1MBD	153	Phillips (1980)	3641	1·99

* Filename in Brookhaven Protein Data Bank.

† Number of residues in the asymmetric unit.

‡ From equation (15).

§ From equation (17).

Table 2. *Histogram model parameters* (*at* 4 Å)

t*	v^0	q^0	α†	t*	v^0	q^0	α†
−0·450	0·12	−0·01	201·2	0·465	2·64	−0·01	10·6
−0·285	0·83	−0·05	48·9	0·495	2·55	−0·01	9·6
−0·255	1·31	−0·08	34·0	0·525	2·59	−0·05	7·7
−0·225	1·84	−0·08	35·5	0·555	2·31	−0·01	12·9
−0·195	2·37	−0·05	58·6	0·585	2·26	−0·02	8·9
−0·165	2·64	0·09	114·7	0·615	2·28	−0·04	6·4
−0·135	2·03	0·49	175·5	0·645	2·13	−0·03	6·5
−0·105	−1·04	1·48	166·5	0·675	2·02	−0·03	11·4
−0·075	−7·58	3·28	32·3	0·705	1·84	−0·01	6·3
−0·045	−16·41	5·56	68·0	0·735	1·92	−0·04	6·0
−0·015	−22·23	7·05	210·9	0·765	1·82	−0·04	7·9
0·015	−21·04	6·78	238·0	0·795	1·58	−0·01	9·3
−0·045	−13·84	4·99	66·4	0·825	1·52	−0·01	9·3
0·075	−4·93	2·72	35·6	0·855	1·38	−0·00	6·4
0·105	0·94	1·15	44·4	0·885	1·30	−0·00	8·7
0·135	3·55	0·34	54·9	0·915	1·15	0·00	7·4
0·165	4·26	0·04	31·0	0·945	1·02	0·00	8·3
0·195	4·10	−0·02	17·1	0·975	0·81	0·02	10·7
0·225	3·87	−0·03	12·3	1·005	0·69	0·02	19·4
0·255	3·67	−0·03	12·2	1·035	0·55	0·02	18·5
0·285	3·46	−0·02	11·5	1·065	0·34	0·04	19·8
0·315	3·20	−0·00	18·5	1·095	0·29	0·03	38·1
0·345	3·15	−0·02	16·0	1·125	0·19	0·02	36·4
0·375	2·98	−0·01	22·2	1·155	0·15	0·01	64·9
0·405	3·06	−0·04	11·2	1·185	0·06	0·02	57·6
0·435	2·86	−0·03	11·9	1·5	0·00	0·00	191·2

* The values of t correspond to the middles of bins.

† Model accuracy characteristics α are defined in equation (16).

where

$$n_k^{(j)} = N_{\rm tot}^{(j)} \nu_k^{(j)} \Delta_k = (N_{\rm tot}^{(j)} F_{000}^{(j)} \Delta_k / V^{(j)}) v_k^{(j)},$$

$$n_{k,{\rm theor}}^{(j)} = N_{\rm tot}^{(j)} \nu_{k,{\rm theor}}^{(j)} \Delta_k = (N_{\rm tot}^{(j)} F_{000}^{(j)} \Delta_k / V^{(j)}) v_{k,{\rm theor}}^{(j)}.$$

Conditional minimization (13)-(14) was performed by the Lagrange multipliers method.

2.4. *Choosing the criterion of histogram closeness*

When determining $\{v_k^0\}_{k=1}^K$ and $\{q_k^0\}_{k=1}^K$ from the minimum discrepancy condition (13) at 4 Å resolution, we have a value of Q equal to $0·36 \times 10^5$. Table 1 gives the values of simulated and 'real' histogram discrepancy for base proteins:

$$Q^{(j)} = \sum_{k=1}^{K} (n_{k,{\rm theor}}^{(j)} - n_k^{(j)})^2 / n_{k,{\rm theor}}^{(j)}. \qquad (15)$$

Large values of criteria Q and $Q^{(j)}$ show that the quantity $n_k^{(j)}$, or $n_{k,{\rm theor}}^{(j)}$, is not a proper estimate for the mean-square deviation of 'real' $n_k^{(j)}$ from 'theoretical' $n_{k,{\rm theor}}^{(j)}$. The deviation is due, first of all, to a high idealization of (7), rather than to the $n_k^{(j)}$

spreading statistically about their means. When improving structure factors by minimization of (3), it is convenient to choose the weight coefficients w_k so that they reflect the accuracy of the predicted frequencies ν_k. We defined the correction factors to be

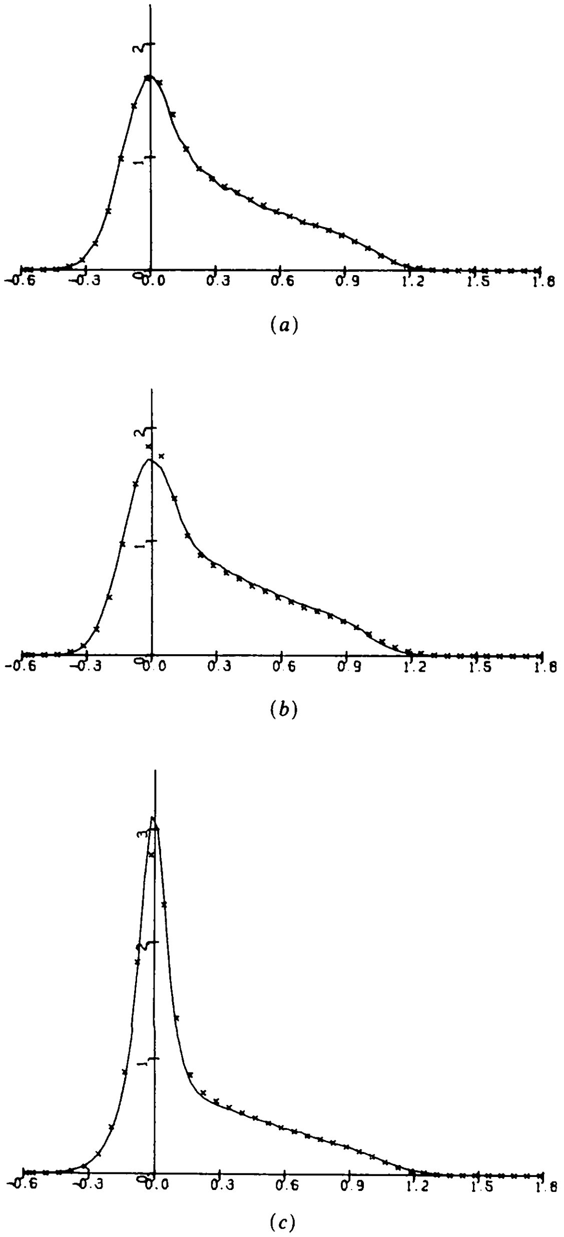
$$\alpha_k = (1/J) \sum_{j=1}^{J} (n_{k,\text{theor}}^{(j)} - n_k^{(j)})^2 n_{k,\text{theor}}^{(j)} \qquad (16)$$

and redefined the criterion of closeness between theoretical and calculated histograms as

$$\begin{aligned} Q_g &= K^{-1} \sum_{k=1}^{K} (n_{k,\text{theor}} - n_{k,\text{calc}})^2 / \alpha_k n_{k,\text{theor}} \\ &= K^{-1} \sum_{k=1}^{K} (N\Delta_k/\alpha_k) \\ &\quad \times (\nu_{k,\text{theor}} - \nu_{k,\text{calc}})^2 / \nu_{k,\text{theor}}. \end{aligned} \qquad (17)$$

Then the mean value of Q_g for base proteins equals one, and it is not reasonable to continue the minimization of criterion (3) when Q_g has reached 1·0 or a lower value.

The values of $\{v_k^0\}_{k=1}^K$, $\{q_k^0\}_{k=1}^K$ and correction factors α_k are given in Table 2. Table 1 gives the values of criteria (15) and (17) for the base proteins. Fig. 5 shows best and worst agreements between simulated and 'real' histograms for base proteins and an example of histogram prediction for a protein absent in the base set.



(a)

(b)

(c)

Fig. 5. The simulated histograms (×××) compared with their 'real' counterparts (——): (a) HIPIP (best agreement among the base proteins); (b) myoglobin (worst agreement among the base proteins); (c) concanavalin (not included in the base protein list).

3. Application: restoring structure factors for γ-crystallin IIIb

We tested histograms simulated by (7) with γ-crystallin IIIb. This is a protein from calf's eye lens. One molecule weighs about 20 000 daltons. The crystals belong to space group $P2_12_12_1$ with unit-cell parameters $a = 58{\cdot}7$, $b = 69{\cdot}5$ and $c = 116{\cdot}9$ Å (Chirgadze, Sergeev, Fomenkova & Oreshin, 1981). The intensity array was collected at a resolution up to 2·5 Å (Chirgadze *et al.*, 1986); this was the resolution at which the structure of γ-crystallin IIIb was refined. Another diffraction set of ('dried') protein crystals was collected at a resolution up to 1·9 Å (Chirgadze *et al.*, 1990), and had a smaller unit cell: $a = 57{\cdot}38$, $b = 70{\cdot}13$, $c = 115{\cdot}4$ Å. For the different modifications the discrepancy in the data over the 2·5 Å area was defined to be

$$R = 2 \sum_s |F_{\text{wet}} - F_{\text{dry}}| \Big/ \sum_s |F_{\text{wet}} + F_{\text{dry}}|,$$

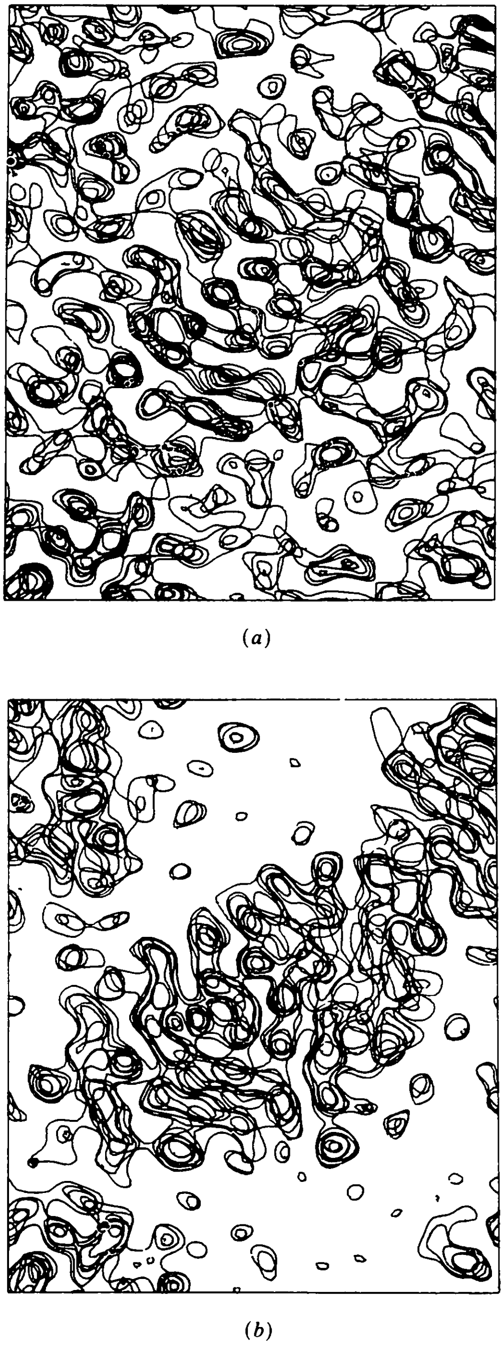
and was equal to 0·255. On technical grounds, the initial 'dry' set lacked considerably many reflections (the area up to 4 Å showed only 2834 of 4224 possible reflections). We used these data in tests aimed at restoring the missing structure factors.

We started with a 4 Å synthesis over 2852 reflections (of 4224 in the independent part of the unit cell) with the coefficients $F_{\text{dry}}(\mathbf{s}) \exp[i\varphi_{\text{wet}}(\mathbf{s})]$, where $F_{\text{dry}}(\mathbf{s})$ were the amplitudes of the second, 'dry', modification, and $\varphi_{\text{wet}}(\mathbf{s})$ were the phases calculated from the atomic model for the 'wet' protein modification. Sections of the synthesis are shown in Fig. 6(*a*).

The synthesis was of poor quality because it contained about two thirds of the reflections only, and its phases contained errors, being of the first modification, not the second.

Then we tried to determine the missing structure factors (both phases and moduli) through the condition of the minimum of criterion (17).

Theoretical frequencies were determined from the model (6)-(7), whose coefficients $\{v^0_k\}^K_{k=1}$, $\{q^0_k\}^K_{k=1}$ were found from the base protein set (Table 1). The set does not include γ-crystallin. Fig. 7 shows the theoretical histogram $\{\nu^0_k\}^K_{k=1}$, which corresponds to the starting synthesis of Fig. 6(*a*) and the histogram which represents the resulting synthesis with restored structure factors. Its sections are shown in Fig. 6(*b*).


(*a*)

(*b*)

Fig. 6. (*a*) Sections $z = 17/240$-$23/240$ for the 4 Å start synthesis for the dry form of γ-crystallin IIIb; (*b*) the improved maps. (Lowest contours on both figures bound 30% of the unit-cell volume.)

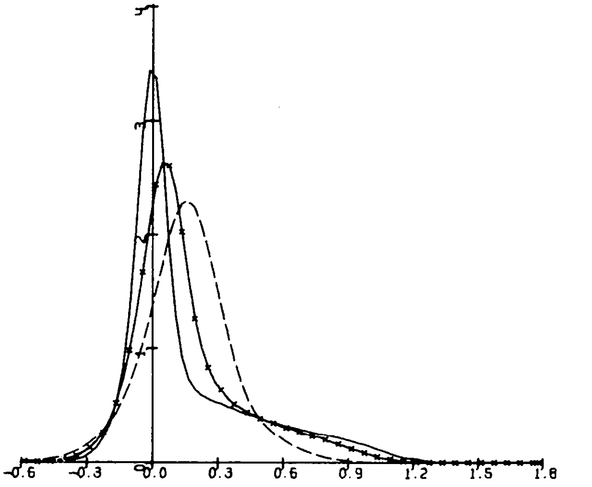


Fig. 7. Simulated (——), start (— —) and final (—×—) histograms for the dry form of γ-crystallin IIIb.

The authors thank A. G. Murzin and A. G. Urzhumtsev for valuable discussions and O. M. Liguinchenko for her help in preparing the manuscript.

References

Artymiuk, P. J. & Blake, C. C. F. (1981). *J. Mol. Biol.* **152**, 737-762.

Blake, C. C. F., Geisow, M. J., Oatley, S. J., Rerat, B. & Rerat, C. (1978). *J. Mol. Biol.* **121** 339-356.

Borkakoti, N., Moss, D. S. & Palmer, R. A. (1982). *Acta Cryst.* **B38**, 2210-2217.

Chirgadze, Yu. N., Nevskaya, N. A., Fomenkova, N. P., Nikonov, S. V., Sergeev, Yu. V., Brazhnikov, E. V., Garber, M. B., Lunin, V. Yu., Urzhumtsev, A. G. & Vernoslova, E. A. (1986). *Dokl. Acad. Nauk SSSR*, **290**, 2, 492-495.

Chirgadze, Yu. N., Nevskaya, N. A., Vernoslova, E. A., Nikonov, S. V., Sergeev, Yu. V., Brazhnikov, E. V., Fomenkova, N. P., Lunin, V. Yu. & Urzhumtsev, A. G. (1990). *Exp. Eye Res.* In the press.

Chirgadze, Yu. N., Sergeev, Yu. V., Fomenkova, N. P. & Oreshin, V. D. (1981). *FEBS Lett.* **131**, 81-84.

Cotton, F. A., Hazen, E. E. & Legg, M. J. (1979). *Proc. Natl Acad. Sci. USA*, **76**, 2551-2555.

Dijkstra, B. W., Kalk, K. H., Hol, W. G. J. & Drenth, J. (1981). *J. Mol. Biol.* **147**, 97-124.

Dodson, E. J., Dodson, G. G., Hodgkin, D. C. & Reynolds, C. D. (1979). *Can. J. Biochem.* 469-475.

Freer, S. T., Alden, R. A., Carter, C. W. & Kraut, J. (1975). *J. Biol. Chem.* **250**, 46-54.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T. & Xuong, Ng. H. (1970). *Biochemistry*, **9**, 1997-2009.

Furey, W., Wang, B. C., Yoo, C. S. & Sax, M. (1983). *J. Mol. Biol.* **167**, 661-692.

Guss, J. M. & Freeman, H. C. (1983). *J. Mol. Biol.* **169**, 521-563.

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

HAŠEK, J. (1984). *Acta Cryst.* **A40**, 340-346.

HAŠEK, J. & SCHENK, H. (1988). *Acta Cryst.* **A44**, 482-485.

HAŠEK, J., SCHENK, H., RIERS, C. TH. & SCHGEN, J. D. (1985). *Acta Cryst.* **A41**, 333-340.

KANNAN, K. K., NOTSTRAND, B., FRIDBOURG, K., LOVGREN, S., OHLSSON, A. & PETEF, M. (1975). *Proc. Natl Acad. Sci. USA*, **72**, 51-55.

KŘÍŽ, V. (1989). *Acta Cryst.* **A45**, 456-463.

LUNIN, V. YU. (1986). *Use of the Information on Electron Density Distribution in Proteins.* Preprint. Pushchino, USSR.

LUNIN, V. YU. (1988). *Acta Cryst.* **A44**, 144-150.

LUNIN, V. YU., URZHUMTSEV, A. G. & SKOVORODA, T.P. (1990). *Acta Cryst.* **A46**, 540-544.

LUNIN, V. YU. & VERNOSLOVA, E. A. (1990). *Acta Cryst.* Submitted.

LUZZATI, V., MARIANI, P. & DELACROIX, H. (1988). *Makromol. Chem. Macromol. Symp.* **15**, 1-17.

MARIANI, P., LUZZATI, V. & DELACROIX, H. (1988). *J. Mol. Biol.* **204**, 165-189.

MATHEWS, F. S., LEVINE, M. & ARGOS, P. (1971). *Nature (London) New Biol.* **233**, 15-18.

PAPAMOKOS, E., WEBER, E., BODE, W., HUBER, R., EMPIE, M. W., KATO, I. & LASKOWSKI, M. (1982). *J. Mol. Biol.* **158**, 515-538.

PHILLIPS, S. E. V. (1980). *J. Mol. Biol.* **142**, 531-534.

PODJARNY, A. D. & JONATH, A. (1977). *Acta Cryst.* **A33**, 655-661.

SIELESKI, A. R., HENDRICKSON, W. A., BROUGHTON, C. G., DELBAERE, L. T. J., BRAYER, G. D. & JAMES, M. N. G. (1979). *J. Mol. Biol.* **134**, 781-804.

ZHANG, Y. J. & MAIN, P. (1990). *Acta Cryst.* **A46**, 41-46.

SHORT COMMUNICATIONS

Contributions intended for publication under this heading should be expressly so marked; they should not exceed about 1000 *words; they should be forwarded in the usual way to the appropriate Co-editor; they will be published as speedily as possible.*

Acta Cryst. (1991). A**47**, 52-55

Improvement of the tangent formula by constraints based on additional information.

By JORDI RIUS and CARLES MIRAVITLLES, *Institut de Ciència de Materials (CSIC), carrer Martí Franquès s/n,* 08028 *Barcelona, Spain*

(*Received* 17 *April* 1989; *accepted* 20 *August* 1990)

Abstract

The first part of this communication describes a simple procedure by which the non-centrosymmetric form of the tangent formula is adapted to incorporate the 'centrosymmetry constraint' for centrosymmetric structures, thus allowing refinement of phases uniformly distributed from 0 to 2π to the expected values 0 or 2π. The convergence of the resulting formula is illustrated with two structures. In the second part, a modified tangent formula including the constraint based on the zero points of the Patterson function is derived. To do this, both the Cochran integral $\int_v \rho^3\,dV$ and the sum over all zero points of the Patterson function of ρ^2 are expressed in terms of the phases of the strong E's. The modified tangent formula is then obtained assuming that the difference between the two corresponds to a large positive maximum for the correct phases. Finally, it is shown how the information supplied by the weak E's and by the zero points can be treated in an unified way, so that a combined tangent formula can be derived.

Introduction

As is well known, the integral (Cochran, 1952; Hauptman & Karle, 1953)

$$V^2 \int_v \rho^3(\mathbf{r})\,d\mathbf{r} = \text{large magnitude}, \qquad (1)$$

including the 'positivity criterion' of the electron-density distribution, can be expressed as the sum of the triplets

$$\sum_{\mathbf{h}} \sum_{\mathbf{h}'} \mathbf{E}_{-\mathbf{h}} \mathbf{E}_{\mathbf{h}'} \mathbf{E}_{\mathbf{h}-\mathbf{h}'}. \qquad (2)$$

As shown by Debaerdemaeker, Tate & Woolfson (1985), a way of deriving the tangent formula (Karle & Hauptman, 1956)

$$\varphi(\mathbf{h}) = \text{phase of } \left\{ \sum_{\mathbf{h}'} \mathbf{E}_{\mathbf{h}'} \mathbf{E}_{\mathbf{h}-\mathbf{h}'} \right\} \qquad (3)$$

is to assume that the true phase angles $\varphi(\mathbf{h})$ of the normalized structure factors correspond to a maximum of the double summation (2). However, by refining phases with the tangent formula, it is also possible to reach, besides the correct maximum, false maxima. This may happen for several reasons, such as the effect of space-group symmetry, size of the structure and special features in the atomic positions (Schenk, 1988). By using additional information *e.g.* the weak E's found from the diffraction experiment (Debaerdemaeker, Tate & Woolfson, 1985) or the minimum interatomic separation derived from the atomic size (Rius & Miravitlles, 1989), constraints can be added to (1) in order to reduce the number of such false maxima. In this communication, the constraints based on the centrosymmetry of the electron-density distribution and on the zero points of the Patterson function are investigated.

The centrosymmetry constraint

A special case of the conventional tangent formula results from introducing the centrosymmetry constraint explicitly in (1), *i.e.* in the form of an integral:

$$I = V^2 \int_v \rho(-\mathbf{r})\rho^2(\mathbf{r})\,d\mathbf{r}. \qquad (4)$$

The integrals (4) and (1) will only be equivalent for a

0108-7673/91/010052-04$03.00