Table 3. *Test of the theoretical results for the best centrosymmetric models*

	$(A1, S1)$		$(A2, S1)$		$(A3, S1)$		$(B1, S2)$		$(B2, S2)$		$(B3, S2)$	
R index	$\bar{R}$	$\langle\vert\Delta\mathbf{r}\vert\rangle$	$\bar{R}$	$\langle\vert\Delta\mathbf{r}\vert\rangle$	$\bar{R}$	$\langle\vert\Delta\mathbf{r}\vert\rangle$	$\bar{R}$	$\langle\vert\Delta\mathbf{r}\vert\rangle$	$\bar{R}$	$\langle\vert\Delta\mathbf{r}\vert\rangle$	$\bar{R}$	$\langle\vert\Delta\mathbf{r}\vert\rangle$
$\bar{R}_1(F_t)$	9·0	0·061	26·6	0·179	34·7	0·246	5·9	0·041	12·6	0·084	21·6	0·143
$\bar{R}_1(y_t)$	10·4	0·061	30·0	0·175	39·3	0·245	6·8	0·041	14·2	0·081	24·2	0·138
$\bar{R}_1(I_t)$	11·4	0·066	34·0	0·187	44·7	0·249	6·4	0·036	13·0	0·074	23·5	0·130
$\bar{R}_1(z_t)$	15·9	0·062	46·7	0·180	65·3	0·262	9·1	0·040	18·7	0·071	33·8	0·129
${}_B\bar{R}_1(F_t)$	0·9	0·067	7·3	0·190	12·1	0·253	0·3	0·040	1·4	0·083	4·1	0·140
${}_B\bar{R}_1(y_t)$	1·6	0·063	11·9	0·178	20·8	0·250	0·5	0·036	2·3	0·075	6·7	0·131
${}_B\bar{R}_1(I_t)$	0·6	0·064	5·4	0·177	8·4	0·225	0·2	0·040	0·7	0·068	2·1	0·113
${}_B\bar{R}_1(z_t)$	3·2	0·066	24·8	0·191	53·8	0·283	0·6	0·031	2·7	0·061	10·1	0·124
$\bar{R}_1^f(F_t)$	15·2	0·060	43·0	0·185	52·1	0·256	11·1	0·045	25·0	0·095	39·6	0·164
$\bar{R}_1^f(I_t)$	28·5	0·059	70·9	0·181	83·7	0·250	21·4	0·045	44·2	0·094	65·7	0·160
$\langle\vert\Delta\mathbf{r}\vert\rangle_{\text{est}}$		0·063		0·182		0·252		0·039		0·079		0·137
$\langle\vert\Delta\mathbf{r}\vert\rangle_{\text{true}}$		0·064		0·162		0·227		0·041		0·079		0·131

Note: R is in % and $\langle|\Delta\mathbf{r}|\rangle$ is in Å. $\langle|\Delta\mathbf{r}|\rangle_{\text{est}}$ is the average value in the respective columns.

centrosymmetric structure $S1$ would be the BCM for all these three non-centrosymmetric structures $A1$, $A2$ and $A3$. In each case the structure factor calculated using the known coordinates (of $A1$, $A2$ or $A3$) were taken to correspond to F_N. The structure factors calculated using the known coordinates of structure $S1$ were taken to correspond to F_N^c. The overall values of various R indices for the three cases, namely (i) $(A1, S1)$, (ii) $(A2, S1)$ and (iii) $(A3, S1)$, were computed omitting reflections for which $y_N < 0·3$ $(= y_t)$ and these are given in columns 2, 4 and 6 of Table 3. The theoretically expected values of $\langle|\Delta\mathbf{r}|\rangle$ were then estimated from the respective R values by interpolation using the results in Table 2 and the results thus obtained are given in columns 3, 5 and 7 respectively of Table 3. The average of the $\langle|\Delta\mathbf{r}|\rangle$ values thus obtained from the 'observed' overall values of the R indices are given in the row marked $\langle|\Delta\mathbf{r}|\rangle_{\text{est}}$ under the respective columns. The true values of $\langle|\Delta\mathbf{r}|\rangle$ for the three cases, namely $(A1, S1)$, $(A2, S1)$ and $(A3, S1)$, are given in the last row marked $\langle|\Delta\mathbf{r}|\rangle_{\text{true}}$. A similar procedure was used in the case of the centrosymmetric structure of Hanson & Rohrl (1972) (referred to as structure $S2$) and three non-centrosymmetric structures (called $B1$, $B2$ and $B3$) with $\langle|\Delta\mathbf{r}|\rangle = 0·041$, 0·079 and 0·131 Å, respectively, were generated. The relevant final results obtained for the three cases, namely $(B1, S2)$, $(B2, S2)$ and $(B3, S2)$, are also summarized in Table 3. It is seen from Table 3 that there is reasonably good agreement between the corresponding values of $\langle|\Delta\mathbf{r}|\rangle_{\text{est}}$ and $\langle|\Delta\mathbf{r}|\rangle_{\text{true}}$ in all cases.

One of the authors (NE) thanks the University of Madras for financial assistance.

References

ELANGO, N. & PARTHASARATHY, S. (1990). *Acta Cryst.* **A46**, 495-502.

HANSON, A. W. & ROHRL, M. (1972). *Acta Cryst.* **B28**, 2287-2291.

KROON, J. & KANTERS, J. A. (1973). *Acta Cryst.* **B29**, 1278-1283.

LUZZATI, V. (1953). *Acta Cryst.* **6**, 550-552.

PARTHASARATHY, S. & VELMURUGAN, D. (1981). *Acta Cryst.* **A37**, 472-480.

SWAMINATHAN, P. & SRINIVASAN, R. (1975). *Acta Cryst.* **A31**, 310-318.

Acta Cryst. (1990). **A46**, 507-509

A Formula for Electron Density Histograms for Equal-Atom Structures

BY PETER MAIN

Department of Physics, University of York, Heslington, York YO1 5DD, England

(*Received* 21 *September* 1989; *accepted* 29 *January* 1990)

Abstract

A formula is developed which gives the histogram of electron density values for polypeptide structures. The six parameters of the formula have been evaluated and are given for a range of resolutions from 4·5 to 0·9 Å. The formula may be used in density modification techniques of map improvement for small proteins.

Introduction

Some recent papers (Zhang & Main, 1990*a*, *b*; Main, 1990) have described a method of determination and

0108-7673/90/060507-03$03.00

refinement of phases for protein structures which makes use of the image processing technique of histogram matching. The histogram used is the distribution of electron density values in the unit cell and typical histograms are illustrated in Zhang & Main (1990*a*). Since the true electron density is unknown at this stage of the structure determination, the histogram must be obtained either from the electron density map of a similar but known structure or from a formula. Both methods have been used successfully in this work. This paper describes the development of a suitable formula and gives values for the parameters which allow the density histogram of an equal-atom structure to be predicted for resolutions from 4·5 to 0·9 Å.

The histogram formula

At atomic resolution, where the phase-determination method referred to above works best, the electron density may be assumed to consist of approximately Gaussian atoms on a fairly flat background. The formula which we seek should model both of these regions of the cell. That is, the high density'values should be modelled by the histogram of Gaussian peaks and the low values by the histogram of an assumed randomly distributed background. The latter is easily obtained, since the histogram of randomly distributed values is the well known Gaussian function

$$P(\rho)\,\mathrm{d}\rho \propto \exp\left[-(\rho-\rho_m)^2/2\sigma^2\right]\mathrm{d}\rho, \qquad (1)$$

where $P(\rho)$ is the probability density of the electron density ρ, ρ_m is the mean of the distribution and σ is its standard deviation.

The functional form of the histogram of the Gaussian peaks may be derived as follows. The electron density of a Gaussian atom at a distance r from its centre is given by

$$\rho = \rho_0 \exp(-br^2), \qquad (2)$$

where ρ_0 is the peak density and b is a constant. Since ρ is a monotonically decreasing function of r and is spherically symmetric, the probability of finding any particular value of ρ is proportional to the incremental volume at the corresponding value of r; *i.e.*

$$P(\rho)\,\mathrm{d}r \propto 4\pi r^2\,\mathrm{d}r. \qquad (3)$$

Differentiation of (2) gives

$$\mathrm{d}r = -\mathrm{d}\rho/2br\rho. \qquad (4)$$

The negative sign means that r increases as ρ decreases, but it is the absolute value that must be used in the probability expression. Substitution for dr in (3) gives

$$P(\rho)\,\mathrm{d}\rho \propto 2(\pi r/b\rho)\,\mathrm{d}\rho. \qquad (5)$$

Eliminating r between (2) and (5) and leaving out the constant of proportionality, we obtain the expression for the density histogram of a Gaussian atom as

$$P(\rho)\,\mathrm{d}\rho \propto (1/\rho)\ln^{1/2}(\rho_0/\rho)\,\mathrm{d}\rho. \qquad (6)$$

This expression cannot be normalized until the volume containing the atom is known. Note that when $\rho > \rho_0$, (6) gives a meaningless value for the probability.

We now have the two components of the histogram formula - expressions (1) for the low densities and (6) for the high. The change over between the two presents a problem, however, because nowhere is there a natural match of values. A rather clumsy method has therefore had to be adopted, but it works satisfactorily. Two threshold values of ρ are designated as ρ_1 and ρ_2 ($\rho_1 < \rho_2$) such that (1) is used when $\rho < \rho_1$ and (6) is used when $\rho > \rho_2$. When $\rho_1 < \rho < \rho_2$, a cubic polynomial is used which matches (1) at ρ_1 and (6) at ρ_2. The complete formula can be expressed mathematically as:

for $\rho < \rho_1$

$$P(\rho) = N\exp\left[-(\rho-\rho_m)^2/2\sigma^2\right]; \qquad (7a)$$

for $\rho_1 < \rho < \rho_2$

$$P(\rho) = N(a\rho^3 + b\rho^2 + c\rho + d); \qquad (7b)$$

for $\rho_2 < \rho < \rho_0$

$$P(\rho) = N(A/\rho)\ln^{1/2}(\rho_0/\rho). \qquad (7c)$$

The parameter A gives the relative weight of the terms in (7*a*) and (7*c*) and N is a normalizing factor.

At low resolution, the parameters a, b, c, d of the cubic (7*b*) are calculated by matching the function values and the gradients at ρ_1 and ρ_2. Let us define the function values as

$$\begin{aligned} y_1 &= \exp\left[-(\rho_1-\rho_m)^2/2\sigma^2\right]; \\ y_2 &= (A/\rho_2)\ln^{1/2}(\rho_0/\rho_2) \end{aligned} \qquad (8a)$$

and the gradients as

$$\begin{aligned} s_1 &= -\left[(\rho_1-\rho_m)/\sigma^2\right]\exp\left[-(\rho_1-\rho_m)^2/2\sigma^2\right]; \\ s_2 &= -(A/\rho_2^2)[\ln^{1/2}(\rho_0/\rho_2) \\ &\quad + (\rho_0/2\rho_2)\ln^{-1/2}(\rho_0/\rho_2)]. \end{aligned} \qquad (8b)$$

It is also convenient to define the mean slope between the two points as

$$s_0 = (y_1 - y_2)/(\rho_1 - \rho_2). \qquad (8c)$$

The cubic parameters are then obtained from the relationships:

$$a = (s_1 + s_2 - 2s_0)/(\rho_1 - \rho_2)^2; \qquad (9a)$$

$$b = (s_0 - s_2)/(\rho_1 - \rho_2) - (\rho_1 + 2\rho_2)a; \qquad (9b)$$

$$c = s_2 - 3a\rho_2^2 - 2b\rho_2; \qquad (9c)$$

$$d = y_2 - a\rho_2^3 - b\rho_2^2 - c\rho_2. \qquad (9d)$$

Unfortunately, at high resolution, the cubic sometimes exhibits maxima and minima within the useful range of ρ, which is unreasonable for an electron density histogram. This is avoided by removing the requirement that the gradients should match at ρ_1 and replacing it by the condition that there be no maximum or minimum on the curve, *i.e.* that $b^2-3ac=0$. This changes (9*a*) to

$$a=\frac{2s_0+s_2-[3s_2(4s_0-s_2)]^{1/2}}{2(\rho_1-\rho_2)^2} \tag{10}$$

and the parameters *b*, *c* and *d* are evaluated from the relationships (9) as before.

Evaluation of the parameters

The parameters in the histogram formula to be determined are ρ_m, σ, A, ρ_1, ρ_2 and ρ_0. Since the normalizing factor does not alter the shape of the function and as it depends upon the actual application, it is not evaluated at this stage.

The parameter ρ_0 (the maximum value of ρ in the map) was most easily evaluated by inspecting the electron density maps of a number of equal-atom structures. In order to evaluate the other parameters, histograms of known polypeptide structures were calculated and added together, then a curve was fitted to the accumulated histogram with the parameters as variables. In setting up the histograms, care was taken to use only the molecular density and to leave out the surrounding solvent regions. This was achieved by determining the molecular envelope from the known atomic positions and using density only from within the envelope.

The histogram is also a function of resolution, so each histogram was set up at a number of resolutions and the parameters evaluated for each. The low-resolution histograms (4·5 to 2·2 Å) made use of a cubic curve that matched the other curves in value and gradient at ρ_1 and ρ_2 [(7*b*) and (9)]. However, at 2·2 Å this cubic exhibited a small maximum and minimum which became larger as the resolution increased. The higher-resolution histograms (2·2 to 0·9 Å) therefore used the alternative cubic which does not allow maxima or minima and whose parameters are calculated from (10) and (9*b*), (9*c*) and (9*d*).

The method of curve fitting adopted was the simplex algorithm of Nelder & Mead (1964). It is particularly suited to this kind of problem where the function cannot be differentiated with respect to all the parameters and not all parameters are well determined by the data. Electron density values were weighted to perform the fitting, since high density (corresponding to the atomic peaks) is clearly more important than low. The weights used were proportional to $\rho+5$.

Table 1. *Values of the histogram parameters for equal-atom structures*

Equation (9) is used to calculate the cubic parameters.

Resolution (Å)	ρ_m	$2\sigma^2$	A	ρ_1	ρ_2	ρ_0
4·5	0·119	0·065	0·396	0·167	0·845	1·25
3·5	0·036	0·071	0·378	0·098	1·200	1·52
2·8	−0·028	0·084	0·322	0·059	1·243	1·93
2·2	−0·042	0·114	0·225	0·122	1·283	2·73

Table 2. *Values of the histogram parameters for equal-atom structures*

Equation (10) is used to evaluate *a*. The remaining cubic parameters are obtained from (9).

Resolution (Å)	ρ_m	$2\sigma^2$	A	ρ_1	ρ_2	ρ_0
2·2	−0·035	0·121	0·230	0·275	1·382	2·73
1·8	−0·022	0·134	0·139	0·395	1·540	3·64
1·4	0·010	0·142	0·074	0·535	1·843	5·50
1·1	0·011	0·211	0·039	0·375	1·382	8·09
0·9	0·029	0·176	0·020	0·356	1·463	11·45

The results obtained are set out in Tables 1 and 2. The resolutions were chosen such that the number of reflexions approximately doubles with each increase in resolution. The assumptions used in setting up the formula - Gaussian atoms on a fairly flat background - become invalid at low resolution but, surprisingly, the formula still gives a good fit to real histograms at all resolutions tested, *i.e.* between 4·5 and 0·9 Å. At all resolutions, the model curves deviate less from the accumulated histograms to which they were fitted than the differences between the histograms of different structures.

Use of the formula

The histogram formula is used to predict the distribution of electron density values in the histogram matching technique of Zhang & Main (1990*a*) and its later development (Main, 1990; Zhang & Main, 1990*b*). It may be applied at any resolution from 4·5 to 0·9 Å and not just at the values shown in the tables. Values of the histogram parameters at resolutions which are intermediate to those shown in the tables may be obtained with sufficient accuracy by linear interpolation.

The author is grateful to the Rigaku Corporation of Japan for financial support and to the SERC for research grant GR/E89643.

References

MAIN, P. (1990). *Acta Cryst.* **A46**, 372-377.
NELDER, J. A. & MEAD, R. (1964). *Comput. J.* **7**, 308-313.
ZHANG, K. Y. J. & MAIN, P. (1990*a*). *Acta Cryst.* **A46**, 41-46.
ZHANG, K. Y. J. & MAIN, P. (1990*b*). *Acta Cryst.* **A46**, 377-381.