15.1.2.1.3. *The solvent-flattening procedure*

Once the envelope has been determined, solvent flattening is performed by simply setting the density in the solvent region to the expected value, $\rho_{\rm solv}$:

$$\rho_{\rm mod}(\mathbf{x}) = \begin{cases} \rho(\mathbf{x}), & \rho_{\rm ave}(\mathbf{x}) > \rho_{\rm cut} \\ \rho_{\rm solv}, & \rho_{\rm ave}(\mathbf{x}) < \rho_{\rm cut} \end{cases}. \quad (15.1.2.6)$$

If the electron density has not been calculated on an absolute scale, the solvent density may be set to its mean value.

A related method is solvent flipping, developed by Abrahams & Leslie (1996). In this approach, the flattening operation is modified by the introduction of a relaxation factor, γ, where γ is positive, effectively 'flipping' the density in the solvent region.

$$\rho_{\rm mod}(\mathbf{x}) = \begin{cases} \rho(\mathbf{x}), & \rho_{\rm ave}(\mathbf{x}) > \rho_{\rm cut} \\ \rho_{\rm solv} - [\gamma/(1-\gamma)][\rho(\mathbf{x}) - \rho_{\rm solv}], & \rho_{\rm ave}(\mathbf{x}) < \rho_{\rm cut} \end{cases}. \quad (15.1.2.7)$$

The effect of this modification is to correct for the problem of independence in phase combination and is discussed in Section 15.1.4.3.

15.1.2.2. *Histogram matching*

Histogram matching seeks to bring the distribution of electron-density values of a map to that of an ideal map. The density histogram of a map is the probability distribution of electron-density values. It provides a global description of the appearance of the map, and all spatial information is discarded. The comparison of the histogram for a given map with that expected for an ideal map can serve as a measure of quality. Furthermore, the initial map can be improved by adjusting density values in a systematic way to make its histogram match the ideal histogram.

15.1.2.2.1. *Introduction*

Histogram matching is a standard technique in image processing. It is aimed at bringing the density distribution of an image to an ideal distribution, thereby improving the image quality. The first attempt at modifying the electron-density distribution was that by Hoppe & Gassman (1968), who proposed the '3-2' rule. The electron density was first normalized to a maximum of 1 and modified by imposing positivity. Subsequently, the electron density was modified by $\rho_{\rm mod} = 3\rho^2 - 2\rho^3$. Podjarny & Yonath (1977) used the skewness of the density histogram as a measure of quality of the modified map. Harrison (1988) used a Gaussian function as the ideal histogram in his histogram-specification method for protein phase refinement and extension. The choice of the Gaussian function as the ideal electron-density distribution was based on theoretical arguments instead of experimental evaluation. The Gaussian function was also made independent of resolution. Lunin (1988) used the electron-density distribution to retrieve the values of low-angle structure factors whose amplitudes had not been measured during an X-ray experiment. The electron-density distribution was thought to be structure specific and was derived from a homologous structure. Moreover, the histogram was derived from the entire unit cell, including both the protein and the solvent. Zhang & Main (1988) systematically examined the electron-density histogram of several proteins and found that the ideal density histogram is dependent on resolution, the overall temperature factor and the phase error. It is, however, independent of structural conformation. The sensitivity to phase error suggests that the density histogram could be used for phase improvement. The structural conformation independence made it possible to predict the ideal histogram for unknown structures.

15.1.2.2.2. *The prediction of the ideal histogram*

Polypeptide structures in particular, and biological macromolecules in general, display a broadly similar atomic composition, and the way in which these atoms bond together is also conserved across a wide range of structures. These similarities between different protein structures can be used to predict the ideal histogram even when positional information for individual atoms is not available in a map. If the positional information is removed from an electron-density map, then what remains is an unlabelled list of density values. This list is the histogram of the electron-density distribution, which is independent of the relative disposition of these densities. The shape of the histogram is primarily based on the presence of atoms and their characteristic distances from each other. This is true for all polypeptide structures.

The frequency distribution, $P(\rho)$, of electron-density values in a map can be constructed by sampling the map and counting the density values in different ranges. In practice, once the electron-density map has been sampled on a discrete grid, this frequency distribution becomes a histogram, but for convenience, it is treated here as a continuous distribution.

At resolutions of better than 6.0 Å and after exclusion of the solvent region, the frequency distribution of electron-density values for protein density over a wide range of proteins varies only with resolution and overall temperature factor to a good approximation. If the overall temperature factor is artificially adjusted, for example, by sharpening to $B_{\rm overall} = 0$, then the frequency distributions may be treated as a function of resolution only. Therefore, once a good approximation to the molecular envelope is known, the frequency distribution of electron densities in the protein region as a function of resolution may be assumed to be known. Therefore, the ideal density histogram for an unknown map at a given resolution can be taken from any known structure at the same resolution (Zhang & Main, 1988, 1990*a*).

The ideal electron-density histogram can also be predicted by an analytical formula (Lunin & Skovoroda, 1991; Main, 1990*a*). The method adopted by Main (1990*a*) represents the density histogram by components that correspond to three types of electron density in the map. The first component is the region of overlapping densities, which can be represented by a randomly distributed background noise. The second component is the region of partially overlapping densities. The third component is the region of non-overlapping atomic peaks, which can be represented by a Gaussian.

The histogram for the overlapping part of the density can be represented by a Gaussian distribution,

$$P_o(\rho) = N \exp\left[-(\rho - \overline{\rho})^2/2\sigma^2\right], \quad (15.1.2.8)$$

where $\overline{\rho}$ is the mean density and σ is the standard deviation. The region of partially overlapping densities can be modelled by a cubic polynomial function,

$$P_{po}(\rho) = N\left(a\rho^3 + b\rho^2 + c\rho + d\right). \quad (15.1.2.9)$$

The histogram for the non-overlapping part of the density can be derived analytically from a Gaussian atom,

$$P_{no}(\rho) = N(A/\rho)[\ln(\rho_0/\rho)]^{1/2}, \quad (15.1.2.10)$$

where ρ_0 is the maximum density, N is a normalizing factor and A is the relative weight of the terms between equation (15.1.2.8) and equation (15.1.2.10).

If we use two threshold values, ρ_1 and ρ_2, to divide the three density regions, the complete formula can be expressed as

$$P(\rho) = \begin{cases} N \exp\left[-(\rho - \overline{\rho})^2/2\sigma^2\right] & \text{for} \quad 2\rho \le \rho_2 \\ N(a\rho^3 + b\rho^2 + c\rho + d) & \text{for} \quad 2\rho_2 < \rho \le \rho_1. \\ N(A/\rho)[\ln(\rho_0/\rho)]^{1/2} & \text{for} \quad 2\rho_1 < \rho \le \rho_0 \end{cases} \tag{15.1.2.11}$$

The parameters a, b, c, d in the cubic polynomial are calculated by matching function values and gradients at ρ_1 and ρ_2. The parameters in the histogram formula, $\overline{\rho}$, σ, A, ρ_0, ρ_1, ρ_2, can be obtained from histograms of known structures.

15.1.2.2.3. *The process of histogram matching*

Zhang & Main (1990*a*) demonstrated that, at better than 4 Å resolution, the histogram for an MIR map is generally significantly different from the ideal distribution calculated from atomic coordinates. The obvious course is therefore to alter the map in such a way as to make its density histogram equal to the ideal distribution. Unfortunately, there are an infinite number of maps corresponding to any chosen density distribution, so we must choose a systematic method of altering the map.

The conventional method of performing such a modification is to retain the ordering of the density values in the map. The highest point in the original map will be the highest point in the modified map, the second highest points will correspond in the same way, and so on.

Mathematically, this transformation is represented as follows. Let $P(\rho)$ be the current density histogram and $P'(\rho)$ be the desired distribution, normalized such that their sums are equal to 1. The cumulative distribution functions, $N(\rho)$ and $N'(\rho)$, may then be calculated:

$$\begin{aligned} N(\rho) &= \int_{\rho_{\min}}^{\rho} P(\rho)\ \mathrm{d}\rho, \\ N'(\rho') &= \int_{\rho_{\min}}^{\rho'} P'(\rho)\ \mathrm{d}\rho. \end{aligned} \tag{15.1.2.12}$$

The cumulative distribution function of a variable transforms a value chosen from the distribution into a number between 0 and 1, representing the position of that value in an ordered list of values chosen from the distribution.

The transformation may, therefore, be performed in two stages. A density value is taken from the initial distribution and the cumulative distribution function of the initial distribution is applied to obtain the position of that value in the distribution. The inverse of the cumulative distribution function for the desired distribution is applied to this value to obtain the density value for the corresponding point in the desired distribution. Thus, given a density value, ρ, from the initial distribution, the modified value, ρ', is obtained by

$$\rho' = N'^{-1}[N(\rho)]. \tag{15.1.2.13}$$

The distribution of ρ' will then match the desired distribution after the above transformation. The transformation of an electron-density value by this method is illustrated in Fig. 15.1.2.3. The transformation in equation (15.1.2.13) can be achieved through a linear transform represented by

$$\rho'_i = a_i\rho_i + b_i, \tag{15.1.2.14}$$

where $i = \{1, \ldots, n\}$ and n is the number of density bins. The above linear transform is sufficient if the number of density bins is large enough. An n value of about 200 is usually quite satisfactory.

Various properties of the electron density are specified in the density histogram, such as the minimum, maximum and mean density, the density variance, and the entropy of the map. The mean density of the ideal map can be obtained by

$$\overline{\rho} = \int_{\rho_{\min}}^{\rho_{\max}} \rho P(\rho)\ \mathrm{d}\rho. \tag{15.1.2.15}$$

The variance of the density in the ideal map can be obtained by

$$\sigma(\rho) = \left(\overline{\rho^2} - \overline{\rho}^2\right)^{1/2}, \tag{15.1.2.16}$$

where

$$\overline{\rho^2} = \int_{\rho_{\min}}^{\rho_{\max}} \rho^2 P(\rho)\ \mathrm{d}\rho. \tag{15.1.2.17}$$

The entropy of the ideal map can be calculated by

$$S = -\int_{\rho_{\min}}^{\rho_{\max}} P(\rho)\rho \ln(\rho)\ \mathrm{d}\rho. \tag{15.1.2.18}$$

Therefore, the process of histogram matching applies a minimum and a maximum value to the electron density, imposes the correct mean and variance, and defines the entropy of the new map. The order of electron-density values remains unchanged after histogram matching.

Histogram matching is complementary to solvent flattening since it is applied to the protein region of a map, whereas

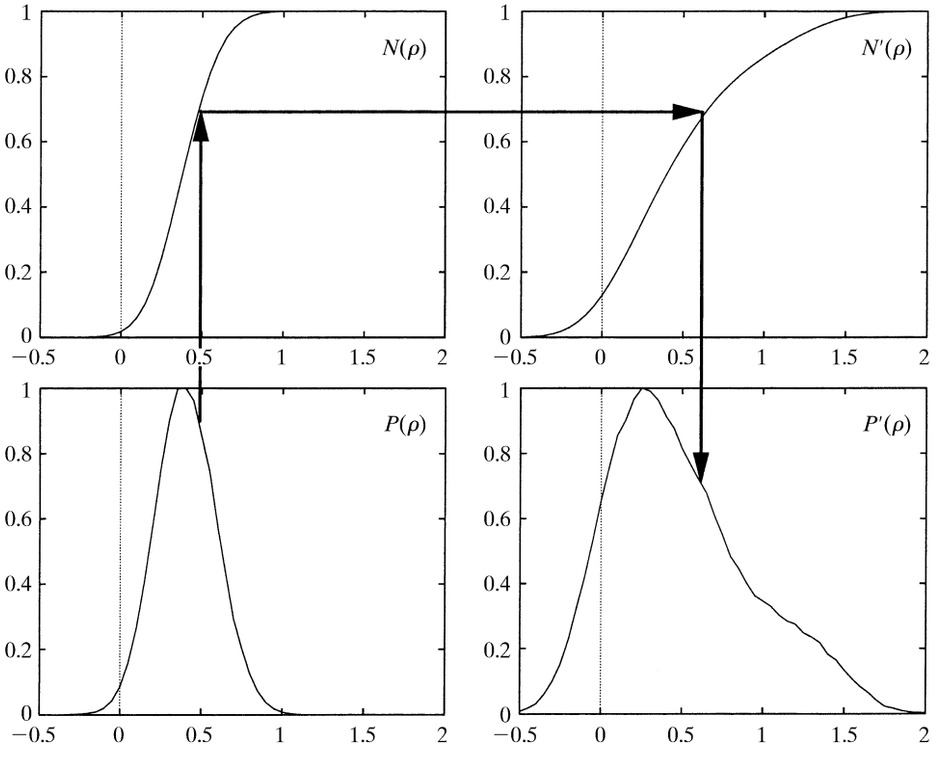

Fig. 15.1.2.3. Transformation of density ρ to $\rho'_{\rm mod}$ by histogram matching.

references