BUTTERWORTH HEINEMANN

Methods for displaying macromolecular structural uncertainty: Application to the globins

Russ B. Altman,* Christopher Hughes,* and Mark B. Gerstein†

**Section on Medical Informatics, Stanford University, Stanford, California*
†Department of Structural Biology, Beckman Center for Structural Biology, Stanford University, Stanford, California

Most molecular graphics programs ignore any uncertainty in the atomic coordinates being displayed. Structures are displayed in terms of perfect points, spheres, and lines with no uncertainty. However, all experimental methods for defining structures, and many methods for predicting and comparing structures, associate uncertainties with each atomic coordinate. We have developed graphical representations that highlight these uncertainties. These representations are encapsulated in a new interactive display program, PROTEAND. PROTEAND represents structural uncertainty in three ways: (1) The traditional way: The program shows a collection of structures as superposed and overlapped stick-figure models. (2) Ellipsoids: At each atom position, the program shows an ellipsoid derived from a three-dimensional Gaussian model of uncertainty. This probabilistic model provides additional information about the relationship between atoms that can be displayed as a correlation matrix. (3) Rigid-body volumes: Using clouds of dots, the program can show the range of rigid-body motion of selected substructures, such as individual α helices. We illustrate the utility of these display modalities by the applying PROTEAND to the globin family of proteins, and show that certain types of structural variation are best illustrated with different methods of display.

Keywords: molecular models, molecular graphics, protein conformation, uncertainty, variation, globins, ellipsoids, Gaussian distributions, correlation matrices, conformational flexibility, macromolecular shape, segmental flexibility, temperature factors

Color Plates for this article are on pp. 190–192.

Address reprint requests to Dr. Altman at the Stanford Section on Medical Informatics, Stanford University, MSOB X-215, Stanford, California 94305-5479.

Received 20 July 1994; revised 3 January 1995; accepted 5 January 1995

INTRODUCTION

The main sources of structural information on biological macromolecules is experimental techniques such as X-ray crystallography and nuclear magnetic resonance (NMR).[1,2] There is also, however, an increasing ability to define structure (or partial structure) with predictive methodologies as well (secondary structure,[3,4] structural class,[5] surface vs. buried atoms,[6] local conformation,[7] approximate fold,[8] and others). Both the experimental and predictive technologies are expected to result in an explosive increase in the amount of structural information available to the biomedical community in the next decade. The Protein Data Bank estimates that the number of new protein entries by the year 2 000 will be 30 000/year.[9] This increase stems primarily from the ability to solve the structure of many related proteins once a prototypical structure has been solved. It is accentuated by computational technologies, such as molecular dynamics, which produce numerous variations of single structures as their dynamic motions are simulated.[10] Finally, there is an increasing ability to make predictions about structure.

Uncertainty of individual structures

All methods for defining or predicting molecular structure have sources of noise that make the exact position of atoms uncertain. This uncertainty can arise from a number of sources. The experimental techniques used to collect structural data are imperfect. The algorithms used to compute structure from these data are imprecise. At the same time, the molecules under study may form a heterogeneous population composed of multiple different conformations or conformations that are rapidly interconverting. In any case,

Journal of Molecular Graphics 13:142–152, 1995

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00002-N

the structures that are computed from uncertain data or from data drawn from a heterogeneous population will have uncertainty associated with their atomic positions. Estimates of this uncertainty are usually provided along with the atomic coordinates, or can be derived from sets of atomic coordinates.

X-ray crystallography Structures determined using X-ray crystallography routinely include information about the temperature factor, or B factor, for each atom. The crystallographic *B* factor is a scalar estimate of signal attenuation derived from electron density maps.[1] Thermal motion of atoms causes a decrease in the intensity of X-ray crystallographic readings by a factor of $e^{-B/K}$, where $B = 8\pi^2\mu^2/3$, μ represents the mean displacement of the atom, and *K* depends on crystallographic experimental parameters. Atoms that reliably assume the same position within the unit cell have a low *B* factor. On the other hand, atoms that do not reliably assume the same position within the unit cell have a high *B* factor. It is well recognized that the positions determined from X-ray crystallographic studies are averages over a population and over a time interval. Thus, there are two factors that contribute to a high *B* factor. First, the population of structures within the crystal lattice may adopt multiple conformations for a given segment. Second, the conformations assumed by a segment may vary during the period of data collection (minutes to hours). In either case, the electron density signal is attenuated, and the *B* factor represents a measure of how localized an atom is within a crystal structure. Although the crystallographer often has information about anisotropies in the *B* factor (which are treated using a parametric representation similar to that described in the next section), the *B* factor is usually reported assuming spherical symmetry. X-ray crystallographic *B* factors vary, depending on experimental conditions, but are generally (and often much) less than 30 Å^2, which represents a mean displacement of approximately 1.1 Å. It has been shown that, for high-resolution structures, most of the *B* factor is due to thermal motion, and not static (lattice) disorder.[11]

Nuclear magnetic resonance Structures determined by NMR must be constructed from multiple distance measurements.[2] There are, therefore, two sources of uncertainty for these structures. First, the interpretation of the NMR data involves assigning a distance (usually a distance range or distance distribution) for each experimental measurement of atomic proximity. Because the data arise from a population of molecules (and are time averaged), there may be errors in assigning this distance. Second, the method used to reconstruct the structure from distances (often called *embedding*) may not find a single optimal structure, but many related structures that all satisfy the distance constraints equally well.[12] For these reasons, NMR structure determinations often produce a family of structures. The mean displacement of atoms in these structures is a function of the number of distance constraints that are extracted from the experiments, and the precision with which the distances can be assigned.[13,14] Small molecules can be solved with a mean displacement on the order of 0.5 Å or less, whereas large structures may have sections with mean displacement on the order of 2.0 to 3.0 Å. The computation of mean displacement is not trivial, because it requires a superposition of structures, and commonly used structural superposition methods can be biased toward particular structure or distorted by a poor choice of atoms used for the superposition. The problem of superposition is discussed further with respect to structural families (below).

Molecular dynamics and predictive technologies Although molecular dynamics is generally not able to produce structures *de novo*, it is often used to explore the conformation and dynamic properties of macromolecules.[10] These simulations allow a single molecule to adopt multiple conformations derived from the starting conformation, based on a simulation of the molecular forces on the molecule. The results of simulations are sets of thousands or millions of structures that differ from each other in subtle ways. These structures can often be displayed as an animation, but it is often useful to summarize the overall features of the animation with static images. These summaries can be generated by considering a superposition of multiple structures (as for NMR), or by extracting the principal significant motions (by analysis of modes or frequency filtering.[15–19] Alternatively, the degree of motion of an atom can be summarized parametrically in a manner similar to the *B* factor of crystallography, in order to summarize the mean displacement of atoms from their average position.[20]

Technologies for predicting structure are not yet mature, but because they are often based on statistical analysis of known structures they are, by definition, subject to sources of error such as insufficient sample; sample bias; and noisy, pooled data. The uncertainty of structures produced by these technologies often has an uneven distribution: certain regions may be reliably modeled to within 1 or 2 Å, while other regions may be uncertain. Some work has been done in an effort to quantify the three-dimensional uncertainty of predicted structures. We have reported a probabilistic algorithm applied to protein structure and RNA structure that produces an explicit estimate of structural uncertainty for each atom.[8,20]

Uncertainty of structural families

There is another area in which the representation of uncertainty within molecular structure becomes critical: the analysis of families of related structures. A family of structures is usually defined by similarity in both secondary structures and their pattern of association. The uncertainty for families of structures is compound: there are uncertainties within each structure and there are differences between structures that define the acceptable variation within the family. One challenge, when comparing molecules in the same family, is to determine the best way to represent the similarities and differences in their structures. The key issue is in how to establish the correspondence between equivalent residues (or nucleic acid bases) in the primary sequence of the structures of interest. This problem has been approached with multiple alignment techniques applied at both the sequence[21,22] and structural level.[23] Given an alignment, a set of equivalent atoms (i.e., atoms that play the same struc-

tural role in each member of the family) can be defined and used for superposition. It is not clear, however, that all atoms for which there are equivalents in each structure should be used for superposition. For example, two equivalent helices that have a consistent relative geometry may be connected by a short coil of constant length. If the coil takes on different conformations in each structure within a related family, then a superposition that attempts to minimize the root mean squared (RMS) deviation of the helices and coil will distribute the variation (which may be essentially all due to the coil) over all three elements, thus leading to a somewhat misleading representation of the variation in the structures. In such a case, what we may actually want is the structurally invariant "core" regions common to all members of the family.[24–28] We have reported an automatic procedure for identifying cores of low structural variance, starting from sequence alignment.[29]

Once a set of equivalent atoms has been defined, however, the problem of representing and displaying multiple molecules within a family becomes similar to the problem of representing multiple possible conformations of the same molecule. In each case, there is a set of atoms, all of which occur in an individual structure, and for which the individual structures all have different three-dimensional shapes. The chief difference between displaying a set of structures from an aligned family and from a single structural determination is that the magnitude of the uncertainty in position is often larger in the case of the aligned family.

In fact, the distinction between the uncertainty in an individual structure and the uncertainty in a family of structures is somewhat arbitrary. For example, different conformations that are sampled during a molecular dynamics calculation represent the variation in a single structure, but they also produce a set of structural coordinates that defines a family of conformations. Similarly, sets of structural studies on proteins that are related by single point mutations have characteristics of a family, but are derived conceptually from a single protein structure. Indeed, members of a protein family are often derived from a common precursor, and so there is a continuum from multiple conformations for a single instance to multiple instances within a family. The way in which we think about the differences between these conformations may change, depending on the questions we are asking.

In this study, we have chosen to illustrate our representations of the globins, a family of structures with a mean RMS deviation between structures of about 2.0 Å.[30] The globin family allows us to accentuate the capabilities of the representations, but the methods presented can be equally well applied to multiple structures from NMR, molecular dynamics, or crystal structures.

We have focused on three categories of display for molecular uncertainty: overlapping molecular stick models, parametric Gaussian distributions, and substructure accessible volumes. Each of these categories accentuates different aspects of structural variation. They all depend on having certain basic information. We assume that for each ensemble of structures, the following information is available.

1. The equivalencies between atomic positions: Each atom included in the superposition must be associated with an equivalent atom in all other structures. In the case of multiple conformations for a single molecular structure, this is trivial; each atom corresponds with itself in the other structures. In the case of multiple structures within the same family, the problem of aligning all the structures, and of assigning atom equivalencies, may be difficult.
2. A subset of atoms to be used in superposition: Given a set of equivalent atoms and structures, it is necessary to pick some subset of the atoms to be used to bring the structures into the same coordinate system. Frequently, all the atoms are used, and this provides an overview of individual atomic variations. As argued above, it is sometimes useful to define a subset of atoms for superposition, and examine the variation of other atoms with respect to a superposition of the subset. These ideas are illustrated in the application to the globin family discussed below (Application to the Globins).
3. The three-dimensional coordinates of every atom for which an equivalency has been established.
4. For sample substructure accessible volume only: Information, provided by the user, about how to group atoms into meaningful structural units such as helices, strands, or higher-order structural units.

Overlapping molecular models

One of the most common ways to compare related structures is to find a superposition of the structures that minimizes the RMS error between them. This has been the subject of careful study, and a number of algorithms have been proposed for finding unbiased superpositions of N structures.[31–34] Most of these algorithms rely on a basic algorithm for superimposing two structures, x_1 and x_2, so that the root mean squared distance (RMSD) between corresponding atoms is minimized:

$$RMSD = \sqrt{\frac{\sum_i (x_{1i} - \mathbf{R}x_{2i})^2}{N}} \quad (1)$$

where x_{1i} is the position of the ith atom in structure x_1, x_{2i} is position of the ith atom in structure x_2, and $\mathbf{R}$ is the rotation required to superpose the structures. Efficient, robust procedures for calculating this (optimal and unique) superposition have been described.[35–39] Some of these methods are based on the singular value decomposition of a specially created tensor. Grid search methods have also been described. The RMSD is a convenient measure of structural similarity, but it is not perfect. For example, it is difficult to relate the same RMSD measure when applied to pairs of structures with different numbers of atoms. In addition, the RMSD does not capture local similarities when they are in a context of global dissimilarity. Nevertheless, for structures that are known to be similar (in both size and overall topology) the RMSD is a reasonable measure of similarity.

Extending the formalism of two structure superposition, the methods for superposing an ensemble of N structures seek to find the rotation matrices for each structure so that

they can be transformed to the same coordinate system.[31] These methods minimize the sum of squared distances between all equivalent atoms:

$$E(\Omega) = \sum_{j<k}^{N} \sum_{i=1}^{M} (\mathbf{R}_j x_{ji} - \mathbf{R}_k x_{ki})^2 \qquad (2)$$

where the first sum is over all pairs, j, k of the N structures in the ensemble Ω, the second sum is over the M aligned positions in each structure, and $\mathbf{R}_j x_{ji}$ represents the rotated coordinates of atom i in structure x_j.

If we rotate each structure using the optimal rotation matrix, $\mathbf{R}_j$, then they can be drawn together in the same coordinate system (in an overlapping fashion) to give an impression of where there are regions of high variation and where there are regions of low variation, as shown in Color Plate 1. This display technique has drawbacks, however, that may make it less useful.

1. As the number of atoms in the structures increases, it becomes difficult to reliably establish the correspondence between equivalent atoms in different structures.
2. As more structures are superimposed, it becomes more and more difficult to examine particular areas of interest within the superimposed structures, because the density of line segments becomes too great.

Of course, the choice of which subset of atoms to use when calculating the rotation matrices in Eq. (2) can drastically affect the quality of the resulting overlap image. For example, if all atoms are used in defining the optimal rotations, then the average variation of all the atoms will be minimized. However, if a subset of atoms within a "core" region (perhaps a centrally located secondary structure or a collection of critical secondary structures) is used to define the best superposition, a different set of apparent variations may result (as shown in Color Plate 1).

METHODS

To address some of the problems with simple overlapping molecular models, we have developed a parametric representation of structure.[20] This representation is based on modeling the distribution of atomic positions as a three-dimensional Gaussian probability distribution. If we have a set of structures, superimposed using some RMS criteria, then we can calculate a mean position and three-dimensional variance and covariance for each atomic coordinate. The mean position can be calculated in a straightforward manner by averaging the coordinates for corresponding atoms after they have been transformed into a common coordinate system. It can be represented as a vector, $\mathbf{x}$, for N atoms:

$$\mathbf{x} = [x_1\ y_1\ z_1\ x_2\ y_2\ z_2\ \ldots\ x_N\ y_N\ z_N] \qquad (3)$$

We can also calculate the variances and covariances of each of these coordinates and place them in a covariance matrix. The diagonal elements of such a matrix contain the variances of each coordinate, and the off-diagonal elements of the matrix contain the covariances between parameters.* The covariance matrix is a $3N \times 3N$ matrix:

$$\mathbf{C}(\mathbf{x}) = \begin{pmatrix} \sigma^2_{x_1} & \sigma_{x_1y_1} & \sigma_{x_1z_1} & \cdots & \sigma_{x_1z_M} \\ \cdot & \sigma^2_{y_1} & & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \cdot & & & & \cdot \\ \sigma_{z_Mx_1} & \cdot & \cdot & \cdots & \sigma^2_{z_M} \end{pmatrix} \qquad (7)$$

Taken together, the mean vector, $\mathbf{x}$, and the covariance matrix, $\mathbf{C}(\mathbf{x})$, represent an uncertain model for the location of each atom. $\mathbf{C}(\mathbf{x})$ also contains information about the relationship between coordinates in its off-diagonal elements. In general, two coordinates may have a complicated functional relationship. The covariance is a linearization of this relationship that essentially specifies whether the values of the two parameters increase or decrease together. Although a primitive summary of potentially complicated dependencies, the covariance is often sufficient (especially using iterative techniques to reduce estimation error) for capturing important parameter relationships in a structural model.[40] In fact, a correlation matrix can be calculated by dividing the elements of the covariance matrix by the product of the standard deviations from the corresponding diagonal positions (Figure 1). For the purposes of display in three dimensions, we focus on the variance and covariance of the coordinate of individual atoms. If we imagine the covariance matrix as a matrix of 3×3 matrices,

$$\mathbf{C}(\mathbf{x}) = \begin{pmatrix} \mathbf{C}(x_1x_1) & \mathbf{C}(x_1x_2) & \cdots & \mathbf{C}(x_1x_N) \\ \cdot & \mathbf{C}(x_2x_2) & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & & \cdot \\ \mathbf{C}(x_Nx_1) & \cdot & \cdots & \mathbf{C}(x_Nx_N) \end{pmatrix} \qquad (8)$$

where each submatrix within $\mathbf{C}(\mathbf{x})$ contains the covariances (or variances, along the diagonal) between the individual coordinates of the two atoms, $\mathbf{x}_i$ and $\mathbf{x}_j$,

$$\mathbf{C}(\mathbf{x}_i\mathbf{x}_j) = \begin{pmatrix} \sigma_{x_ix_j} & \sigma_{x_iy_j} & \sigma_{x_iz_j} \\ \sigma_{y_ix_j} & \sigma_{y_iy_j} & \sigma_{y_iz_j} \\ \sigma_{z_ix_j} & \sigma_{z_iy_j} & \sigma_{z_iz_j} \end{pmatrix} \qquad (9)$$

*The variance, covariance, and correlation are defined, respectively, using standard statistical definitions:

$$\sigma^2_x = \frac{\Sigma x^2}{N-1} - \left(\frac{\Sigma x}{N}\right)^2 \qquad (4)$$

$$\sigma_{xy} = \frac{\Sigma xy}{N-1} - \left(\frac{\Sigma x}{N}\right)\left(\frac{\Sigma y}{N}\right) \qquad (5)$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \qquad (6)$$

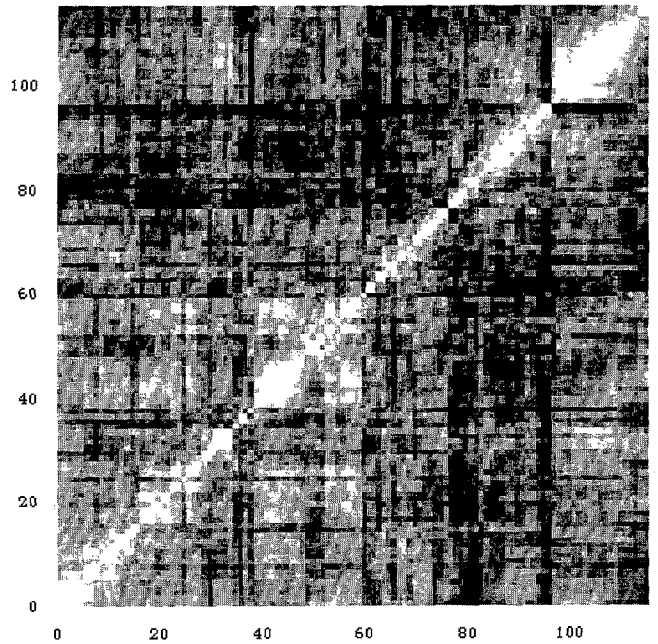

Figure 1. The correlation matrix for the eight globin family members studied. Using the standard alignment described in Ref. 30, we extracted the 115 α-carbon atoms common to the aligned globins. The correlation coefficients between the coordinates of all atoms were calculated using standard definitions. To produce a single scalar summary of the correlations between any two points, each 3 × 3 tensor of correlations between two points was diagonalized, and the sum of the absolute value of the diagonal elements was plotted (thus, perfect correlation or anticorrelation along the principal axes would have a value of 3.0, whereas independent atoms would have a correlation value of 0.0). The resulting 115 × 115 matrix is a measure of the correlations between the 115 atoms, and is displayed here. Low values of correlation are black, and high values are white; shades of gray designate interval values.

the diagonal 3 × 3 matrices are the variance and covariance matrices between the Cartesian coordinates for each atom. Each of these matrices is symmetric and positive definite. The diagonals give the variance of each coordinate along the global *x*, *y*, and *z* coordinates, while the off diagonals give the covariance between the coordinates. Each of these matrices contains information about the three-dimensional distribution of positions for each atom; specifically, they provide the second moment (the variance) of the distribution. If we want to display this information then we need to assume some form for the distribution, define a confidence level or contour at which the distribution should be displayed, and define the orientation of the distribution in space.

1. Assume a distribution, given first two moments: In theory, there are an infinite number of spatial distributions with the same mean and variance. For computational convenience, it is useful to assume a normal distribution with the given mean and variance. In addition, the normal distribution is the least biased estimate of an unknown distribution given only its first two moments.[41] In this case, of course, the full distribution is actually known (it is a discrete distribution of points from each structure), but we are seeking a simplification and representation that is useful when there are hundreds of structures.
2. Define a confidence level: Given the assumption of a normal distribution, it is reasonable to draw a contour for each atom that encloses some expected percentage of all positions. Just as a one-dimensional normal distribution truncated at two standard deviations (SD) contains 96% of the data points contained within this distribution, so a three-dimensional normal (represented by the 3 × 3 variance/covariance tensor) defines an ellipsoidal contour at 2 SD that contains 96% of the locations contained within this distribution. (A 1-SD contour would contain 72% of expected points.)
3. Define orientation: If we imagine an ellipsoid centered at the origin and with major and minor axes aligned with Cartesian axes, then we can describe that ellipsoid with a tensor that has the length of each semiaxis in the diagonal elements, and with zero off-diagonal elements. The off-diagonal elements are zero because an aligned ellipsoid has no correlations between coordinates: each coordinate can be selected from a normal distribution with appropriate variance independent of the other coordinates. Now, we can rotate this tensor by an arbitrary three-dimensional rotation matrix and then the off-diagonal elements will be nonzero, because the rotation introduces correlations between coordinates in the global coordinate system. To display our covariance matrices, we reverse this process. In particular, the submatrices on the diagonal of the large covariance matrix may be decomposed, using a Jacobi decomposition,[42] to

$$\mathbf{C} = \mathbf{RDR}^T$$

where **D** is a diagonal tensor whose diagonal elements are the variances of a three-dimensional distribution that is oriented along the global coordinates. **R** contains the eigenvectors of **C**, which define a rotation matrix that transforms points in the coordinate system of **D** into the coordinate system of **C**. The variance of a distribution is the square of its standard deviation, thus if we take the square root of the elements of **D** and multiply by two (if we want a 2-SD contour), then we have the axes of an ellipsoid, oriented along the coordinate axes. We can apply rotation matrix, **R**, to this ellipsoid to produce the correct orientation, and then translate to the mean position (taken from vector **x**), in order to display the three-dimensional distribution most consistent with the sample of points, as shown in Color Plate 2. This method of representing variation in atomic position provides some benefits over the overlap methods.

1. All the structural variation is represented as an ellipsoid that can be easily labeled or turned on/off for quick identification.
2. The complexity of the display does not change even as the number of structures is increased. Every atom has a single ellipsoid, and the only variation with sample size is the accuracy of the mean and variances.

However, the method is still subject to the same sensitivity to and dependence on the choice of atoms for defining the optimal alignment (see Color Plate 2).

It should be noted here that the crystallographic B factor, as given in Eq. (1), is related in a simple way to this probabilistic representation of atomic position. The squared mean displacement is, by definition, the variance of the distance from the mean position. Instead of representing the variance as a three-dimensional tensor, the isotropic B factor assumes a spherically symmetric distribution. Thus, using the notation developed above, the B factor can be inserted into a 3 × 3 tensor as follows:

$$\mathbf{C} = \begin{pmatrix} \frac{3B}{8\pi^2} & 0 & 0 \\ 0 & \frac{3B}{8\pi^2} & 0 \\ 0 & 0 & \frac{3B}{8\pi^2} \end{pmatrix}$$

The ellipsoids then degenerate into spheres, which can be drawn to provide an indication of which sections of the structure have the most thermal motion (see Color Plate 3).

Sampled substructure accessible volumes

The final method we have implemented for examining the uncertainty of a macromolecule is meant primarily for structures for which relatively well-defined subunits can be identified, but for which the relative position of the subunits is variable. For example, we may have a set of proposed structures for a macromolecular complex that has a number of helical elements that occur in different positions. The internal structure of each helix can be modeled in a local coordinate system. Once a global coordinate system is established (perhaps around a single helical element), then the position of the local coordinate systems of each other helix can be described as a translation and a rotation from the origin of the global coordinate system. For compactness, we store the position of the local coordinate system as three Cartesian coordinates, and the orientation of the object as three Euler angles.[43]

To illustrate the spatial variation in the position of these subunits, we can systematically disperse dots within the volume of the structural element in the local coordinate system, and then transform all of the dots to the actual position in the global coordinate system and render them. As all the possible locations for each structural element are drawn in this fashion, a cloud of these dots is formed, which gives the viewer a feel for the overall spatial extent of possible locations for the structural element. These clouds can be color coded in order to label positions for different structural elements. This is illustrated in Color Plate 4. This representation is particularly useful for defining sets of relative positions between substructures. By fixing one object in the global coordinate system and drawing the accessible volume for the other objects, we can describe the range of positions they occupy with respect to the fixed object.

APPLICATION TO THE GLOBINS

We have written a software package, PROTEAND, that is designed to facilitate the use of the representations described in the previous section. PROTEAND is written in C, uses the GL graphics library from Silicon Graphics, Inc. (as well as the X and MOTIF window systems), and is available by anonymous ftp on camis.stanford.edu, /pub/altman/proteand.tar. PROTEAND is a display program only, and does not perform the analyses required for these modes of display (programs to do these analyses are available on request from authors). It does, however, use a set of general file formats that are flexible and easy to use once these analyses are complete. The program can draw structures using standard stick figures, generalized ellipsoids, or combinations of spheres and cylinders (as shown in Color Plates 1–5). Among other features, the program allows for real-time rotation of all images (using wireframe ghosts). It gives user control of all object colors, drawing styles, background colors, object reflectivity, light source position, depth-cueing parameters, and other display parameters. It can also save commonly used combinations of these parameters (including orientation and zoom level) in order to facilitate instant reloading of favorite views. To underscore the different strengths and weaknesses of these display modalities, we illustrate them with the same example: the core regions of the globin molecules. From the Protein Data Bank,[44] we chose eight structures from the globin family that have been the subject of previous investigations[29,30,45]: 1ECD, 1MBA, 1MBD, 2HBG, 2LH4, 2LHB, and the A and B chains of 3HHB. (All structures are of the deoxy form except for 1MBA and 2LHB.) We used the canonical numbering scheme used by Lesk and Chothia in aligning these structures manually.[30] The eight structures each have eight helices (A through H), seven of which are included in the alignment of Lesk and Chothia, and which are shown in Color Plate 5. All analyses were done at the α-carbon level, because the identity of many residues changes over these eight structures. The following analyses were performed.

1. Overlapping structures: Using all 115 atoms in the alignment of Lesk and Chothia, we calculated an unbiased average structure, using a method described in Ref. 29, and yielding results identical to those produced using the method of Diamond.[31] We then fit all eight structures to this average, and this superposition is displayed in Color Plate 1A. To illustrate the sensitivity of these methods to choice of criteria for alignment, we also fit all 8 structures to the average of a subset of 28 atoms (of the initial 115) that define helices A and B, whose relative positions are invariant over the 8 helices as described in Ref. 29, and to a subset of 13 α carbons that belong to a helix in close contact with the heme moiety (helix F), which has relatively high variance in position compared to the other 7 helices. These are displayed as overlapping line drawings in Color Plates 1B and C.
2. Parametric probabilistic representation: Using the structures that were fitted for the overlap display, we calculated means and variances as described in Methods. We used a 2-SD cutoff for the ellipsoids. Color Plate 2 shows the ellipsoids of uncertainty for the globins, corresponding to the same superposition strategies described for the overlapping structures (i.e., all atoms, helices A and B, and helix F alone).
3. PROTEAND is able to display PDB files directly, using

the *B* factor to scale the volume of each atom. We selected and displayed each α carbon of 2HHB (subunit A) with a ball equal in size to its mean displacement. Atoms with *B* factors above 25 Å^2 were colored red (see Color Plate 3).

4. Figure 1 shows a graphical representation of the correlation matrix [as defined in Eqs. (7) and (8)] for the 8 globin structures (using all 115 atoms in the standard alignment[30]). Using an *ad hoc* measure to summarize the overall correlation between the coordinates of all pairs of atoms (as described in the caption to Figure 1), we checked the three-dimensional positions of the pairs of atoms that had the highest correlation that were not part of the same helix. These are shown in Table 1.
5. Substructure accessible volume: Each of the globin folds included in the ensemble has eight helical elements (A–H). For each helix position within each protein (again, using the same alignment as in Color Plates 1A and 2A), we transformed into the common coordinate system a set of dots scattered within the cylinder defined by the backbone atoms. This defines a cloud of possible locations for each helix in the globin family, and is shown in Color Plate 4.

Table 1. Highest correlations between atoms that are not part of the same helix[a]

Position 1	Position 2	Correlation	Distance (Å)
E20	E-F1	2.79	3.72
B5	E4	2.71	11.61
A10	F4	2.71	24.78
B15	A17	2.70	23.51
B6	E4	2.70	8.66
B14	E19	2.67	24.05
B5	E5	2.66	9.37
F9	F-G1	2.65	3.60
B8	E18	2.65	19.90
C6	H14	2.65	25.72
C3	H13	2.65	21.79
B14	E15	2.65	18.61
C4	H13	2.65	19.36
A11	F4	2.64	23.69
A9	E4	2.64	25.13
B14	E8	2.63	12.02
B14	E20	2.62	27.12
B12	E19	2.62	21.64
F-G4	G1	2.61	3.72
C6	H12	2.61	23.07
B14	E6	2.61	15.50
B15	E19	2.60	26.03
C6	H13	2.60	24.01
B10	E4	2.60	7.53

[a]The first column gives the position of the α carbon in the standard numbering scheme.[30] The second column provides the same information for a second atom. The third column reports the sum of the magnitudes of the diagonalized correlation matrix, and the final column shows the average distance of these atoms from one another in the eight structures. High coordinate correlation within this group of 115 aligned α carbons does not imply physical proximity.

DISCUSSION

The importance of good structural models in biology cannot be overestimated. At the molecular level, models are used for drug design or functional analysis. It is critically important that application programs have access to both the certainty with which the structure is known, as well as the important correlations and covariances between individual structural parameters. A static molecular structure may be less useful for the process of drug design than a structure in which the regions of structural uncertainty are clearly defined. Sometimes it is possible to determine the cause for structural uncertainty: NMR measurements can sometimes indicate which regions are mobile and thus provide an explanation for lack of strong signals. Frequently, however, the exact source of uncertainty is not clear.

PROTEAND is designed to complement the existing body of graphics software used to display macromolecules.[46–53] The program complements interactive graphics programs, such as Insight,[49] which have many features to display and analyze structures, but which have no features to represent the uncertainty in an ensemble of structures. It may be particularly appropriate for analyzing the results of an NMR structure determination. Analyzing uncertainty in crystallographic structures is a somewhat different situation. The crystallographic model-building programs FRODO[46] and O[47,48] represent the uncertainty in a single structure in terms of contours of electron density. This is obviously the best representation for the uncertainty in the real data when building models. However, for the noncrystallographer, electron density is difficult to display and hard to interpret. PROTEAND can summarize a large amount of electron density information using its ellipsoid atom representation. The ellipsoidal representation is similar to that implemented in the ORTEP program[53a] for anisotropic *B* factors, but takes advantage of modern hardware rendering capabilities. PROTEAND makes use of the dot-cloud and helix-cylinder representations popularized in earlier programs. Like the dot-surface programs of Connolly,[53] PROTEAND uses semitransparent clouds of dots to represent features of the protein structures while still allowing the atomic skeleton to be visible. However, unlike the Connolly programs and GRASP (Nicholl et al.[52]), PROTEAND uses dots to represent backbone atoms rather than surfaces. In the substructure accessible volume representation, PROTEAND uses the cylinders popularized by Lesk and Hardmann[50] to represent helices.

Sensitivity to superposition criteria

Color Plates 1, 2, and 4 accentuate the importance of selecting reasonable overlap criteria. All helices in the globin fold are not equal. There are a number of lines of evidence that show that helices A and B, and the end of G near them, have relatively fixed spatial relationships with each other, while others are much more variable from globin to globin. NMR evidence shows that helices D and F are most mobile in solution and the last to fold.[54] In addition, experimental folding analyses suggest that helix F assumes an unusual and potentially less stable geometry relative to the rest of the molecule.[55] Subbiah et al. have shown that helices A,

B, C, G, and H can be aligned with low residual errors with helices from the helix-turn-helix repressor family, suggesting that these constitute a core fold that has been reused during evolution.[26] We have reported a core finding procedure that identifies these same helices as less structurally variable than the others.[29] It may, therefore, be preferable to use segments with low structural variability for superimposing the globins. To illustrate the effects of choosing different subsets of common segments for superposition, we performed a superposition of all the structures, based on an optimal superposition of helix F and helices A and B, in addition to using all segments. Helices A and B are the least variable helices across the globin family. Helix F is the most variable helix with respect to the core helices. Color Plates 1C and 2C show clearly that, when helix F is used to define the optimal superposition, the uncertainty for the globin helices is extremely large. From the perspective of helix F, there is a large amount of variation in the positions of all other helices within the globin fold. It may be that the subtle variations of helix F in concert with the heme group account for the different functional characteristics of the globins.

Directional variation clear from probabilistic representation

It is clear in Color Plate 2B that the uncertainty ellipsoids for helix E are oriented in the same direction as the axis of the helix. Closer examination reveals that the orientation of the helix axis does not vary greatly over the eight structures, but there is a significant difference in the register or phase of the helix, as some helices (especially the one from 2HGB, which is yellow in Color Plate 1) are shifted upward (or downward) along the long axis in comparison with others. Although it may be equally easy to infer this pattern from the overlapping stick-figure representation, the probabilistic image has an underlying mathematical representation of this variation that makes it easily accessible to automated recognition and analysis: the covariance matrix for each atom, when translated to the local coordinate system of the helix, will show a consistently larger variation in the direction of the helical long axis. This is in contrast to the situation for helix B, in which the ellipsoids are not oriented along this axis but reflect a more uniformly distributed uncertainty.

The volumes of the ellipsoids drawn in Color Plate 2 are quantitative measures of the uncertainty in position for the substituent atoms. The average volumes of the ellipsoids can be used as an indication of the amount of uncertainty in the position of each helix. We have found that, when all atoms are used for the superposition, helix F has the largest average uncertainty followed (in descending order) by helices E, G, C, H, A, and B. Therefore, helix B seems to be the most structurally conserved in terms of its position relative to other helices.[29]

Interpretation of correlation matrix

The correlation matrix highlights atoms whose motions are the most correlated. With a small sample size (only eight structures included in this calculation), we must interpret these cautiously. Nevertheless, high values of correlation may indicate significant dependencies. We found that atoms in the same helix tended to have high summed correlation coefficients (often above 2.9), which reflects the fact that these helices tend to move as units. Table 1 lists pairs of atoms that are highly correlated but not in the same helices. It is somewhat surprising that these correlations do not imply physical proximity: amino acids 10 and 67 are separated by more than 25 Å on average in the eight globin structures, but have an average correlation coefficient in each of the three principal directions above 0.9. The intriguing possibility that these correlations are important for functional reasons is a subject of current investigation.†

Fulcrum variation is clear from substructure accessible volume

Unlike the parametric probabilistic view, the substructure sampled volumes do not shed light on the individual variations of atomic position. Instead, they are useful for viewing the variations in position of entire substructures (helices in this case). For example, examination of Color Plate 4 shows that the positions of helix E and A are distributed as if they were rotating around a fulcrum (going through the helix, perpendicular to the long axis). In contrast, helix B seems to have a volume of uncertainty more compatible with lateral translations of the helical long axis and not rotation around a fulcrum. Differences in substructure orientation are difficult to extract from overlap images and individual uncertainty ellipsoids. However, the mode of display in which entire substructures are drawn to produce a cloud of possible locations makes such differences quite evident.

Relative "motion" of different substructures

By choosing different subsets of atoms to use for superposition, we essentially redefine the global coordinate system. In the case in which we use the F helix alone, we obtain images that show how the other helices vary from the perspective of helix F. Similarly, when we use the A and B helices we see the minimal ellipsoids for atoms within these helices, and the relative movements of other helices. Of course, this is a family of different structures, and therefore there is no real motion between individual structures, but rather a set of changes in packing and orientation over the family of members. However, these images leave the impression that there is a range of relative positions into which these eight structures (and their substructures) fall. Membership in the globin family, then, might be measured by

†To study the correlation matrix more carefully, one must evaluate the sensitivity of the correlation values to the choice of atoms used for superposition. In addition, when summarizing the correlation between two atomic positions, a variety of measures can be used. As described in the caption to Figure 1, we have summed the absolute values of the diagonalized correlation matrix between two points. This measure has the advantage of taking on the value of 3.0 when the atoms are moving in any combination of perfect correlation and anticorrelation along the coordinate axes, and taking on the value of 0.0 when the atoms are perfectly uncorrelated. This measure, however, is not invariant to rotation of the ensemble. Other methods for summarizing correlation (such as the determinant or the trace of the correlation matrix) emphasize different aspects of correlation.

the degree to which a new structure falls within the bounds defined by previously identified family members.

Crystallographic uncertainty versus family variation

It is clear that the structural imprecision of individual crystal structures (as shown in Color Plate 3) is not the same as structural variation of equivalent atoms across a family of structures (as shown, e.g., in Color Plate 2). The crystallographic *B* factor is determined in part by whether an atom is on the surface, and by the nature of the intermolecular contacts with the crystal lattice. In the three structures shown in Color Plate 3, the only regions that have consistently high *B* factors are the regions between helices C and E, as well as the end of helix A. On the other hand, helix F is the most variable segment across different members of the globin family (as seen in Color Plates 1, 2, and 4). These observations underscore the importance of distinguishing the precision with which the atomic positions are determined and the concordance of these positions in related structures.

Representing uncertainty in large complexes

A final advantage of using substructure representations, such as cylinders for helices, is that they can be used to represent the structures (and structural uncertainty) of larger complexes, because they simplify the image. In the case of the globins, by focusing on the positions of eight cylinders, we can better understand the relationship between the cylinders and the primary directions in which their positions are uncertain. We have also used PROTEAND to display images from structure calculations involving 10 RNA helices and 5 proteins. By using simplified representations, we are able to present these large macromolecular ensembles in a manner that is easier to interpret.[56]

CONCLUSIONS

The anticipated explosion in the availability of structures produced by both experimental and predictive technologies makes the issue of representing and manipulating these structures in a uniform manner critically important. As these structures are made available, we cannot predict the array of uses to which they will be put. The technologies used to define structure are not perfect, and they will produce structures in which the reliability of subsegments varies greatly. It is therefore important to have technologies for representing and manipulating these structures at the proper level of precision. It is unreasonable to expect all investigators to be intimately familiar with the sources of uncertainty from each of the technologies used to define structure, and yet it is important that they be stored in a common location with a common representation. Therefore, representations of structure should capture notions of variability and covariation explicitly, preferably in a technique-independent manner. Then, no matter what the source of the uncertainty, segments of structure can be labeled as uncertain for the users. The users, at the same time, can come to expect a uniformity of representation that makes a detailed understanding of the experimental conditions unnecessary for at least a subset of tasks.

We have developed PROTEAND in part to focus attention on the lessons that can be learned by considering structural variation and structural uncertainty, both within individual structures and across families of structures. Our results with hemoglobin illustrate the kinds of insight these representations can provide. First, we are able to see areas of high and low overall uncertainty. Helix F is the most variable (largest ellipsoids), while helices A and B are the least variable (smallest ellipsoids). The effects of including these helices in superposition criteria can also be gauged by seeing the effect on the ellipsoidal volumes. Second, the shape of the ellipsoids enables us to appreciate the principal directions of uncertainty in atomic position. Most of the uncertainty for helix E is concentrated along the long axis, with relatively low uncertainty in orthogonal directions. Third, secondary structure and domain motions can be visualized clearly by using substructure abstractions in which clouds of dots give an overall impression of the variability of these substructures. The particular range of relationships between substructures can also be represented by fixing the coordinate system around structures of interest and examining the variation in positions of other elements. These conclusions about variation within the globin fold and their possible biological significance are more difficult to draw from the overlap displays that are typically created. The newer methods described here complement the overlap methods, and should become standard parts of molecular display packages.

The long-term goal of this work is to develop a methodology for both representing and manipulating biological structural information, especially with respect to the uncertainty within individual structures, and the variation across related biological structures. This article emphasizes the graphical end points of some of the representations we have developed. However, the mathematical representations underlying these display modalities are useful for primary computation as well, and have been used to solve protein structures using NMR data,[57] and to model RNA structure.[8,56] In addition, we have used representations based on analysis of positional variation for defining key core elements in a family of structures.[29]

ACKNOWLEDGMENTS

R.B.A. is a Culpeper Medical Scholar, and is supported by NIH Grant LM-05652. M.B.G. is supported by a Damon-Ruyon Walter-Winchell fellowship (DRG-1272). Computing environment provided by the CAMIS resource under NIH Grant LM-05305. Parts of PROTEAND have derived from ideas contained within a display program originally written by Craig Cornelius and John Brugge as part of the PROTEAN project at Stanford University.

REFERENCES

1 Blundell, T.L. and Johnson, L.N. *Protein Crystallography*. Academic Press, New York, 1976
2 Wuthrich, K. *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, New York, 1986

3 Stolorz, P., Lapedes, A., and Xia, Y. Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.* 1992, **225**(2), 363–377
4 Gibrat, J.F., Robson, B., and Garnier, J. Influence of the local amino acid sequence upon the zones of the torsional angles phi and psi adopted by residues in proteins. *Biochemistry* 1991, **30**(6), 1578–1586
5 Metfessel, B.A. and Saurugger, P.N. Pattern recognition in the prediction of protein structural class. In *Hawaii International Conference on Systems Science*. IEEE Computer Society Press, Los Alamitos, California, 1993, pp. 679–688
6 Holbrook, S., Muskal, S., and Kim, S. Predicting surface exposure of amino acids from protein sequence. *Prot. Eng.* 1990, **3**(8), 659–665
7 Sippl, M.J. and Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* 1992, **13**(3), 258–271
8 Altman, R.B. Probabilistic structure calculations: A three-dimensional tRNA structure from sequence correlation data. In *First International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California, 1993, pp. 12–20
9 Koetzle, T. and Abola, E. Brookhaven National Laboratories. Personal communication
10 Levitt, M. and Sharon, R. Accurate simulation of protein dynamics in solution. *Proc. Natl. Acad. Sci. U.S.A.* 1988, **85,** 7557–7561
11 Petsko, G.A. and Ringe, D. Fluctuations in protein structure from X-ray diffraction. *Annu. Rev. Biophys. Bioeng.* 1984, **13**(331)
12 Havel, T. and Wuthrich, K. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular ^1H-^1H proximities in solution. *Bull. Math. Biol.* 1984, **46**(4), 673–698
13 Clore, G.M., Robien, M.A., and Gronenborn, A.M. Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.* 1993, **231**(1), 82–102
14 Zhao, D. and Jardetzky, O. An assessment of the precision and accuracy of protein structures determined by NMR. *J. Mol. Biol.* 1994, **239**(5), 601–607
15 Dauber-Osguthorpe, P. and Osguthorpe, D.J. Extraction of the energetics of selected types of motion from molecular. *Biochemistry* 1990, **29**(36), 8223–8228
16 Levitt, M. Real-time interactive frequency filtering of molecular-dynamics trajectories. *J. Mol. Biol.* 1991, **220,** 1–4
17 Daggett, V. and Levitt, M. Realistic simulations of native-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.* 1993, **22,** 353–380
18 Karplus, M. and Petsko, G.A. Molecular dynamics simulations in biology. *Nature (London)* 1990, **347,** 631–639
19 McCammon, J.A. and Harvey, S.C. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, New York, 1987
20 Altman, R.B. *Exclusion Methods for the Determination of Protein Structure from Experimental Data*. Stanford University Dissertation, Stanford, California, 1989
21 Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A., and Wootton, J.C. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science* 1993, **262**(5131), 208–214
22 Haussler, D., Krogh, A., Mian, I.S., and Sjolander, K. Protein modeling using hidden Markov models: Analysis of globins. In *Hawaii International Conference on Systems Science*. IEEE Computer Society Press, Los Alamitos, California, 1993
23 Holm, L. and Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 1993, **233**(1), 123–138
24 Greer, J. Comparative modeling methods: Application to the family of mammalian serine proteases. *Proteins* 1990, **7,** 317–334
25 Taylor, W.R. and Orengo, C.A. Protein structure alignment. *J. Mol. Biol.* 1989, **208,** 1–22
26 Subbiah, S., Laurents, D.V., and Levitt, M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* 1993, **3,** 141–148
27 Lesk, A.M. *Protein Architecture: A Practical Approach*. Oxford, IRL Press, 1991
28 Chothia, C. and Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 1986, **5,** 823–826
29 Altman, R.B. and Gerstein, M.B. Finding an average core structure: application to the globins. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, California, 1994, 19–27
30 Lesk, A.M. and Chothia, C.H. How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol.* 1980, **136,** 225–270
31 Diamond, R.D. On the multiple simultaneous superposition of molecular structures by rigid-body transformations. *Protein Sci.* 1992, **1,** 1279–1287
32 Gerber, P.R. and Müller, K. Superimposing several sets of atomic coordinates. *Acta Crystallogr.* 1987, **A43,** 426–428
33 Kearsley, S.K. An algorithm for the simultaneous superposition of a structural series. *J. Comput. Chem.* 1990, **11,** 1187–1192
34 Shapiro, A. and Botha, J.D. A method for multiple superposition of structures. *Acta Crystallogr.* 1992, **A48,** 11–14
35 Arun, K.S., Huang, T.S., and Blostein, S.D. Least-squares fitting of two 3-D point sets. *IEEE Trans. Pattern Anal. Machine Intell.* 1987, **9**(5), 698–700
36 Diamond, R. A comparison of three recently published methods for superimposing vector sets by pure rotation. *Acta Crystallogr.* 1989, **A45,** 657
37 Lesk, A.M. A toolkit for computational molecular biology. II. On the optimal superposition of two sets of coordinates. *Acta Crystallogr.* 1986, **A42,** 110–113
38 McLachlan, A.D. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr.* 1972, **A28,** 656–657
39 McLachlan, A.D. Rapid comparison of protein structures. *Acta Crystallogr.* 1979, **A38,** 871–873
40 Altman, R.B. A probabilistic algorithm for calculating

structure: Borrowing from simulated annealing. In *Ninth Annual Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Mateo, California, 1993, 23–31
41 Smith, R.C. and Cheeseman, P. On the representation and estimation of spatial uncertainty. *Int. J. Robotics Res.* 1986, **5**(4), 56–68
42 Press, W.H., Flannery, B.P., Teukolsky, S., and Vetterling, W.T. *Numerical Recipes in C*. Cambridge University Press, New York, 1992
43 Altmann, S. *Rotations, Quaternions, and Double Groups*. Clarendon Press, Oxford, 1986
44 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542
45 Bashford, D., Chothia, C., and Lesk, A.M. Determinants of a protein fold: Unique features of the globin amino acid sequences. *J. Mol. Biol.* 1987, **196,** 199–216
46 Jones, T.A. Interactive computer graphics: FRODO. *Methods Enzymol.* 1985, **115,** 157–171
47 Jones, T.A. and Thirup, S. Using known substructures in protein model building and crystallography. *EMBO J.* 1986, **5,** 819–822
48 Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. Improved methods for building protein models in electron density: Maps and the location of errors in these models. *Acta Crystallogr.* 1991, **A47,** 110–119
49 Dayringer, H.E., Tramontano, A., Sprang, S.R., and Fletterick, R.J. Interactive program for visualization and modelling of proteins, nucleic acids and small molecules. *J. Mol. Graphics* 1986, **4,** 82–87
50 Lesk, A.M. and Hardmann, K.D. Computer-generated schematic diagrams of protein structures. *Science* 1982, **216,** 539–540
51 Kraulis, P.J. MOLSCRIPT—A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* 1991, **24,** 946–950
52 Nicholls, A., Sharp, K.A., and Honig, B. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* 1991, **11,** 281–296
53 Connolly, M. Solvent-accessible surfaces of proteins and nucleic acids. *Science* 1983b, **221,** 709–713
53a Johnson, C.K. *ORTEP: A FORTRAN Thermal-Ellipsoid Plotting Program for Crystal Structure Illustrations*. Report No. ORNL-3794. Oak Ridge National Laboratory, Oak Ridge, Tennessee, 1970
54 Cocco, M.J. and Lecomte, J.T.J. Characterization of hydrophobic cores in apomyoglobin: A proton NMR spectroscopy study. *Biochemistry* 1990, **29,** 11067–11072
55 Jennings, P.A. and Wright, P.E. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. *Science* 1993, **262,** 892–896
56 Altman, R.B., Weiser, B. and Noller, H.F. Constraint satisfaction techniques for modeling large complexes: Application to central domain of 16S RNA. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (Altman, Brutlag, Karp, Lathrop, and Searls, Eds.). AAAI Press, Menlo Park, California, 1994, pp. 10–18
57 Arrowsmith, C., Pachter, R., Altman, R., and Jardetzky, O. The solution structures of *E. coli* Trp repressor and Trp aporepressor at an intermediate resolution. *Eur. J. Biochem.* 1991, **202,** 53–66