Distribution Analysis of the Variation of B-Factors of X-ray Crystal Structures: Temperature and Structural Variations in Lysozyme[†,‡]

John E. Wampler

Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia 30602

Received March 27, 1997[⊗]

The *B*-factor (isotropic temperature factor) data for X-ray structures of hen egg-white lysozyme from the study of Young *et al*. (Young, Dewan, Nave, and Tilton *J. Appl. Cryst*. **1993**, *26*, 309−319) potentially contain information about the relative contributions of static and dynamic variation to these factors. The six structures of the protein were obtained at two widely different temperatures (100 and 298 K), with two crystal forms (monoclinic and tetragonal) and other experimental differences. In addition, the monoclinic lysozyme crystals with two molecules per asymmetric unit allow direct examination of variation between structures determined under identical conditions at both temperatures. The *B*-factors from these structures all have complex distribution functions as might be expected considering all of the influences that these values must reflect. The empirical cumulative distribution functions (eCDF's) of these data show that they are representative of complex, multicomponent distributions. Distribution analysis using the DANFIP procedure (Wampler, *Anal. Biochemistry* **1990**, *186*, 209−218) of the data sets reveals that they can be modeled as four to six Gaussian subpopulations, that these subpopulations do not correlate with specific atom types, specific amino acid residues or fixed locations in the structure. While they do seem to correlate with localized groupings of atoms, these grouping vary from structure to structure even within the same crystal under the same conditions. Temperature seems to have a global effect in this case, but it is clear that other factors including experimental error influence the distribution of B-factors within a given structure. This analysis also helps explain the oft observed lack of atomic level correlation between experimental *B*-factors and calculated mean square displacements from molecular dynamics simulations.

INTRODUCTION

One of the most intriguing aspects of the molecular scale is the dynamics of atomic motion and how they impact on molecular function in biological chemistry.[1] Indeed, the foundation of biological chemistry is chemical kinetics and the vast increase in reaction rate catalyzed by enzymes. To large part these rate increases are fundamentally about the molecular vibrations and coordinated atomic motions. Therefore, it is of extreme interest to try and understand the nature of atomic motion in biomolecules in aqueous solution. From the theoretical point-of-view, these issues are being probed by computational models of structure dynamics and by simulations of molecular motion.[1,2] From the experimental point-of-view, the dynamics of motion at the atomic level are probed by a very few techniques such as nuclear magnetic resonance relaxation times,[3] incoherent neutron scattering[4] and X-ray structure *B*-factors.[5] The focus of this work is on the *B*-factors of structures taken from the Protein Databank.[6]

It is clear that *B*-factor (or isotropic temperature factor) data of crystal structures are influenced by a number of disparate variables.[7] From the theoretical point of view, they reflect on the dynamic and static disorder of the molecule and crystal (eq 1).

$$B = 8\pi^2[\langle \mu_d^2 \rangle + \langle \mu_s^2 \rangle] \quad (1)$$

where $\langle \mu_d^2 \rangle$ and $\langle \mu_s^2 \rangle$ are the mean squared displacements of the atom due to dynamic motion and static variation in the crystal respectively. In a globular protein structure the dynamic disorder measured by $\langle \mu_d^2 \rangle$ should also be dependent upon an atom's position in the structure, i.e., tightly packed, core atoms having low values and loosely packed, surface atoms having larger values. There is also the issue of anisotropy in the dynamic motion of the atoms, however, with protein structures an isotropic model is generally used. Even in those few cases where anisotropic temperature factors are available, the PDB entries will contain calculated isotropic values as well. To the extent that it is not accounted for elsewhere, any measure of crystal disorder, $\langle \mu_s^2 \rangle$, also contains information about changes in electron density due to local interactions and polarizability.

The experimental isotropic *B*-factors are determined by an iterative structure solution procedure with essentially four variables per atom (the *x*-, *y*-) and *z*-coordinates and the *B*-factor). Because of the application of more stringent constraints on the position parameters (for discussion see Stroud and Fauman[8]), the *B*-factors can be biased to reflect the effects and errors due to a variety of experimental variables such as differences in absorption of X-rays for different length paths through the crystal, differences in

[†] Keywords: distribution analysis, B-factors, X-ray structure, lysozyme, molecular dynamics, DANFIP, temperature effects.

[‡] Abbreviations and Symbols: DANFIP = distribution analysis by nonlinear fitting of integrated probabilities; eCDF = empirical cummulative distribution function; 3LYTA = designation of one of the structures of lysozyme from the Brookhaven Protein Data Bank file PDB3LYT.ENT; 3LYBB = designation of the other structure of lysozyme from the Brookhaven Protein Data Bank file PDB3LYT.ENT; 4LYTA = designation of one of the structures of lysozyme from the Brookhaven Protein Data Bank file PDB4LYT.ENT; 4LYTB = designation of the other structure of lysozyme from the Brookhaven Protein Data Bank file PDB4LYT.ENT; 5LYT = designation of the structure of lysozyme from the Brookhaven Protein Data Bank file PDB5LYT.ENT; 6LYT = designation of the structure of lysozyme from the Brookhaven Protein Data Bank file PDB6LYT.ENT.

[⊗] Abstract published in *Advance ACS Abstracts,* October 1, 1997.

S0095-2338(97)00225-4 CCC: $14.00

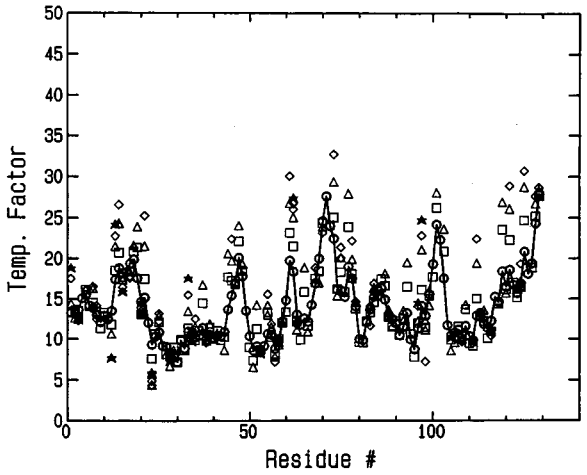

Figure 1. Selected *B*-factors from the tetragonal 298 K structure file (6LYT) are plotted for each residue of hen egg white lysozyme. Symbols: circle, CA *B*-factor; square, CBs; triangle, CG or average of CG1 and CG2; diamond, CD or average of CD1 and CD2; and star, CE or average of CE1 and CE2.

crystal packing, radiation damage and differences in unresolved occupancy of atomic positions. In some cases, specific corrections for some of these effects may be included in the refinement procedure, but not always.

One approach to resolving the various contributions to the experimental *B*-factors has been to study the variation of these factors with temperature.[5,9−11] Another is to study their variation between structures of similar or identical proteins.[8,12,13] The study of Young *et al.*[14] contains information from both perspectives as well as focusing on radiation induced decay. However, the typical analysis in these cases involves use of the standard tools of parametric statistics or examination of the discrete values with no focus on the appropriate statistics and form of the distribution of these values. For example, in many of these studies the arithmetic mean of the entire set of *B*-factors is studied and a Gaussian distribution is assumed. RMS (root-mean-square) averages are also heavily used. However, it is clear from even the most straight forward analysis that *B*-factors and position variations are not distributed as single Gaussian distributions in a structure and, thus, more appropriate statistics must be used in describing their variation.

In spite of the well-known caution against using the standard tools of parametric statistics with data that are not sampled from a single normal or Gaussian distribution or for biased data, there are a number of issues of real data analysis that tend to inhibit more detailed or more appropriate analysis. One limit, for example is how to tell if a small sample of data is indeed normally distributed. Many of the tests routinely described in the literature do not resolve differences with the small samples of real data that are routinely available. For example, one might examine the histogram of the data to see if it approximates the "bell" shape of a normal distribution function or the higher order moments to see if they are consistent with a Gaussian distribution. However, it takes a fairly large set of data before the histogram unambiguously takes the shape of the parent distribution function (see ref 15, Figure 2). Similarly, the higher order moments are "noisy" when used for this purpose (see Wampler[15] Table 1).

If the answer to the question about the nature of the distribution is that it is complex or non-Gaussian, another problem is what practical tools can be used to analyze it. One approach in this case is to use the tools of nonparametric statistics[16,17] to more appropriately parameterize the distribution as a whole. A clear and simple way to represent the central tendency in this case is to use the median value of the data set rather than the arithmetic mean. The breadth of the distribution can be quantified by an interquartile range rather than the standard deviation. However, with real data, the characteristics of the distribution as a whole even with these more robust indicators do not necessarily help when we are trying to understand the complexity of the data or, more importantly, in trying to understand changes in this complexity from experiment to experiment. Distribution analysis by non-linear fitting of integrated probability functions (DANFIP) allows accurate, detailed extraction from data of the number of components and their characteristic parameters for complex distributions.[15] It also provides for assessment of the quality of the model distribution function and a method for correcting truncation error in sampling. One example of its use was reported where single cell pH measurement were collected during the early development of the cellular slime mold *Dictyostelium discoideum*.[18] Assuming that a single normal distribution described the randomness of the measurements gave values that might be considered as typical of a measurement "out-of-control", i.e., large temporal fluctuations in the standard deviation by a factor of about 5. As shown in that work, however, these results could be explained by a model where two discrete normal distributions of values were being sampled and where the critical difference between samples at different times was not in the characteristics of the two parent distributions, which analysis showed were reasonably consistent, but in the relative size of the two populations.

The large database of structural and dynamics data from X-ray crystallography now available[6] is a tremendous resource for examining the variation in structure and dynamics dependant on a wide variety of variables, both experimental, computational and natural. For example, many structures are available of proteins that are homologs or differ in single site mutations. A number of data sets are available of the same protein structure determined under different conditions or by different groups. A wide range of data is available refined with different computational procedures. With regard to *B*-factors, these data allow analysis of such questions as

(1) What is the variation for the same structure determined under identical experimental conditions?

(2) What is the variation with temperature?

(3) What are the influences of other experimental variables?

This analysis of the Young *et al.* data sets[14] addresses these questions in part. In addition it illustrates the value and power of the DANFIP procedure when applied to parameters from X-ray structures.

METHODS

The programs used in this work are available over the internet by anonymous ftp (selene.biochem.uga.edu, directory pub/programs/) or over the World Wide Web (http://selene.biochem.uga.edu/programs.html). Each program file and a descriptive text file are combined in a single compressed file in each case mentioned (extension .ZIP) or

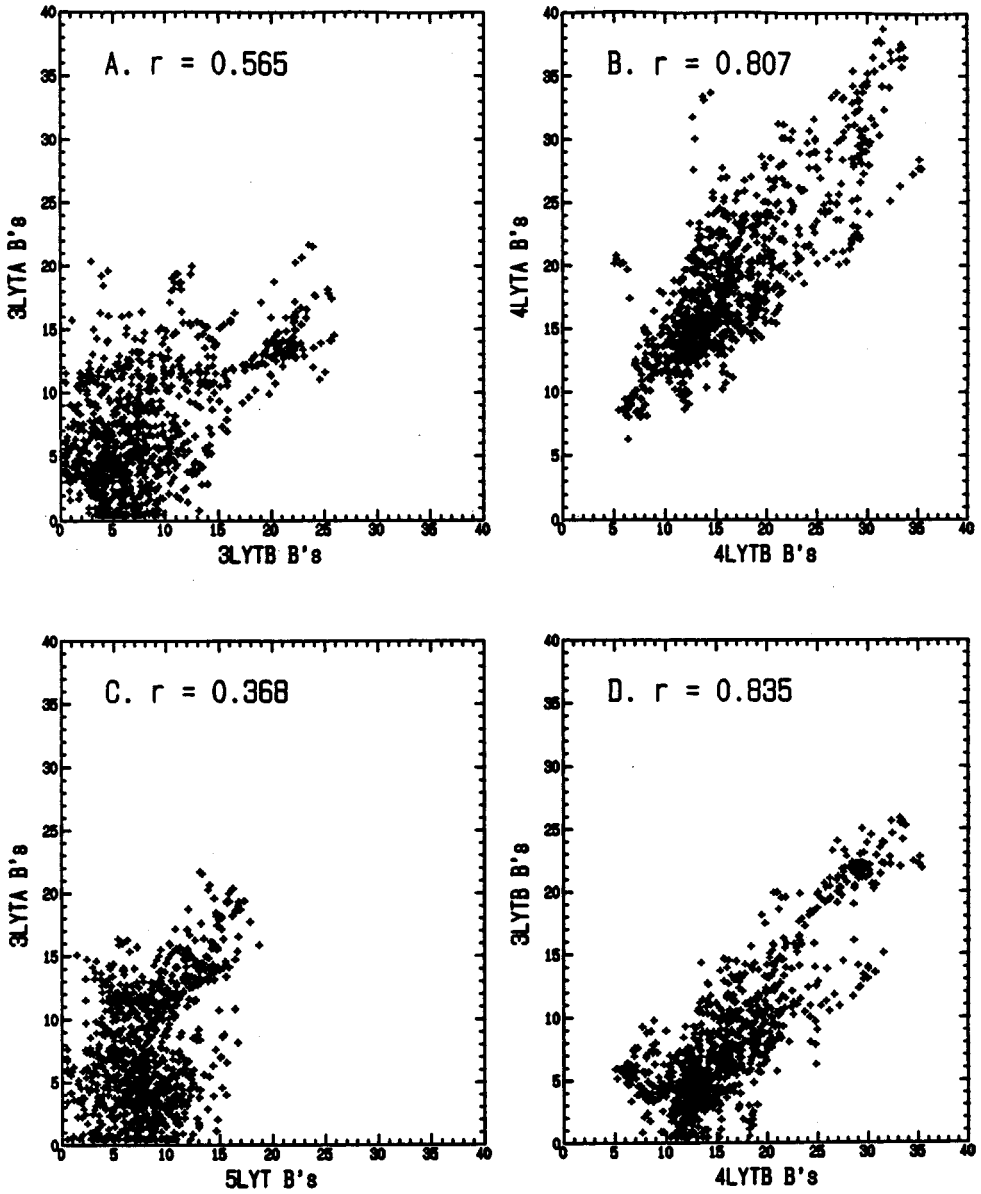

Figure 2. Scatter plots of *B*-factors of corresponding structures (see axis labels) of hen egg white lysozyme. Linear correlation coefficients are given with the panel labels. Panels C and D represent the worst and best case, respectively, for all of the pairwise comparisons of Table 3.

Table 1. Overall Statistics of *B*-Factors (Å^2) for the Lysozyme Structures

file	temp (K)	arithmetic mean	SD	median	interquartile range	skewness	Kurtosis	global mean
3LYTA	100	7.42	4.77	6.59	3.69−11.30	0.48	−0.61	7.36[a]
3LYTB	100	8.41	5.78	6.85	4.32−10.59	1.18	0.72	8.38
5LYT	100	8.12	3.44	7.83	5.68−10.24	0.31	−0.20	8.10
4LYTA	298	18.63	6.02	17.40	14.34−21.81	0.90	0.57	18.63
4LYTB	298	16.46	6.05	15.26	12.36−19.32	0.88	0.41	16.43
6LYT	298	15.16	5.90	13.53	10.69−18.12	1.11	0.89	15.14

[a] For the global mean, truncated data were removed before fitting (see Table 4).

in a self extracting file (extension .EXE). The READ-ME.TXT file on the programs directory gives a fuller description of what is available.

The data extraction and processing methods used in this analysis are implemented in programs written using the Quick BASIC compiler (version 3.0, Microsoft Corporation). The program PDB-GEO.BAS extracts geometric data from Protein Data Bank files. It may be used to extract *B*-factors, atom-to-atom distances, angles and torsion angles. Input to this program can be in the form of a single PDB file or a database file containing the names of multiple PDB files. The output of this program is two types of files, a simple text table of the values and a SPECOS SA compatible data file (see below). Distances, angles and torsions may be extracted from a chain, specific amino acid residues or all residues. The atoms involved do not have to be bonded or in the same residue.

Data processing and graphing was carried out using SPECOS SA, version 3.72, an updated version of the program described earlier.[19] There are a number of added

functions over this previous version which are described in the documentation.

Standard statistical parameters are reported as calculated by SPECOS SA. DANFIP analysis[15] of multiple Gaussian populations were carried out using either SPECOS SA (for 1 and 2 subpopulations) or a special purpose double precision, nongraphics fitter DANFIP.BAS (for three to six subpopulations). The DANFIP procedure uses nonlinear least squares fitting to fit models of multiple overlapping Gaussian distribution functions to the empirical cumulative distribution function (eCDF) created from a sorted set of random data. Therefore, the reduced chi square of the fit gives a criterion for selection between models since it corrects for the improvement in fit that comes from simply increasing the complexity of the model. The corresponding parent population distribution for a fit was determined by differentiation of the fit curve (using SPECOS SA). The components of the fit were also constructed and graphed using SPECOS. Color coding for display was implemented by replacing the actually *B*-factors in copies of the structure file with unitized values (one value for each range to be colored). The images of these structures were then created using the RASMOL structure display program.[20]

The X-ray structure data sets used in this work were obtained from the Protein Data Bank[6] via a local site that mirrors the PDB internet site (http://www.pdb.bnl.gov). The Young *et al*. structures[14] were solved from data taken at two temperatures and with two crystal forms, monoclinic and tetragonal. The monoclinic lysozyme crystals (files 3LYT and 4LYT) had two molecules per asymmetric unit. These structures were analyzed separately as 3LYTA, 3LYTB, 4LYTA and 4LYTB. The tetragonal structures are referred to as 5LYT and 6LYT respectively. 3LYTA, 3LYTB and 5LYT were determined from crystals mounted on glass fibers after removal of excess mother liquor at an experimental temperature of 100 K. 4LYTA, 4LYTB and 6LYT were from crystals in glass capillaries surrounded by mother liquor at 298 K. Other experimental and data processing differences are also described in their paper[14].

RESULTS

The *B*-factor data for these structures clearly do not sample simple normal distributions of such data. The first indication of this and the simplest test is comparison of the medians and arithmetic means (Table 1). With samples this large (1001 values each) from a simple normal distribution, these two values would be nearly identical. These distributions are all slightly skewed to the right (skewness > 0), but have a roughly Gaussian peak shape (kurtosis ∼ 0) (Table 1) considering the magnified variability of these parameters with finite size samples.[15]

One obvious reason that these data do not follow a simple Gaussian distribution is that they combine effects on rigidly held main-chain atoms and more loosely held side-chain atoms. Figure 1, shows the separation of the *B*-factor data for the 298 K tetragonal lysozyme structure (6LYT) according to residue position, i.e., main-chain CA, and side chain positions CB, CG, CD, and CE. For the CG, CD, and CE values, the *B*-factors of any two branched chain atoms (including aromatics) were averaged. For a significant portion of the residues the *B*-factor increases as bonded-distance from the main chain increases. However, even these more narrowly defined data sets are not Gaussian distributed with shape characteristics much like the global data sets (mean versus median, skewness, and kurtosis). In addition, their means and medians show no clear trend with position from main chain (Table 2).

Table 2. Statistics of *B*-Factors of Carbons CA to CE for the Tetragonal Lysozyme Structure at 298 K (6LYT)[a]

carbon position	mean	SD	median	interquartile range	skewness	Kurtosis
CA	14.04	4.27	13.21	10.52−16.71	0.96	0.66
CB	14.35	4.80	13.28	10.47−17.41	0.79	−0.14
CG	15.67	5.81	14.21	11.49−19.32	0.63	−0.34
CD	15.96	6.87	14.09	10.50−20.68	0.67	−0.47
CE	13.01	5.58	11.29	9.72−15.16	1.10	0.41

[a] *B*-factors (Å^2) for all carbons of all residues from the main chain, CA, position to the fourth carbon on the side chain (CE). Data from branched and aromatic residues (e.g., CD1 and CD2) combined.

Table 3. Linear Correlation Coefficients for the *B*-Factors of Lysozyme Structures from the Study of Young *et al.*[14] Based on Atom-to-Atom Comparison

	3LYTA	3LYTB	5LYT	4LYTA	5LYTB	6LYT
3LYTA	1.000	0.565	0.368	0.703	0.604	0.563
3LYTB		1.000	0.464	0.718	0.835	0.524
5LYT			1.000	0.559	0.512	0.636
4LYTA				1.000	0.807	0.733
4LYTB					1.000	0.624
6LYT						1.000

Two approaches might be considered in analyzing these data, one would be to separate the data before analysis into more narrowly defined groupings. At the extreme, each atom could be considered as sampling a unique distribution of its *B*-factor. At the other extreme, the entire data set of all *B*-factors would be analyzed as a sample from a complex, multi-component distribution. In the first case, if the *B*-factors correlate only with each individual atom, then the values from two structures determined under identical conditions might be expected to show a high level of correlation. Figure 2 (panels A and B) shows that this correlation between the two monoclinic 100 and 298 K structures is modest, at best. The linear correlation coefficients for all combinations of two of these structures are given in Table 3. The best and worst of the correlations are shown in Figure 2 (panels C and D). In light of the failure to identify a reasonable degree of correlation with position from the main chain (Table 2) and at the atomic level (Table 3), the prudent approach would seem to be to analyze the combined data.

A clearer picture of the nature of these distribution functions is seen when the empirical cumulative distribution functions, eCDFs, are examined and analyzed. The eCDF is to the integrated distribution function what the more familiar histogram is to the distribution function itself (Figure 3).

The process of constructing the eCDF is illustrated by the panels of Figure 4 where the *B*-factor data for the tetragonal 100 K structure are processed. The raw data (panel A) is first sorted in ascending order (panel B). The eCDF is generated by "swapping" the *X*- and *Y*-axes (panel C, solid line). The dashed line in this panel is the best fit of this data to a single integrated Gaussian function scaled to 1001 (the number of points in the data set). While the general sigmoid shape of the distribution is close to a single Gaussian

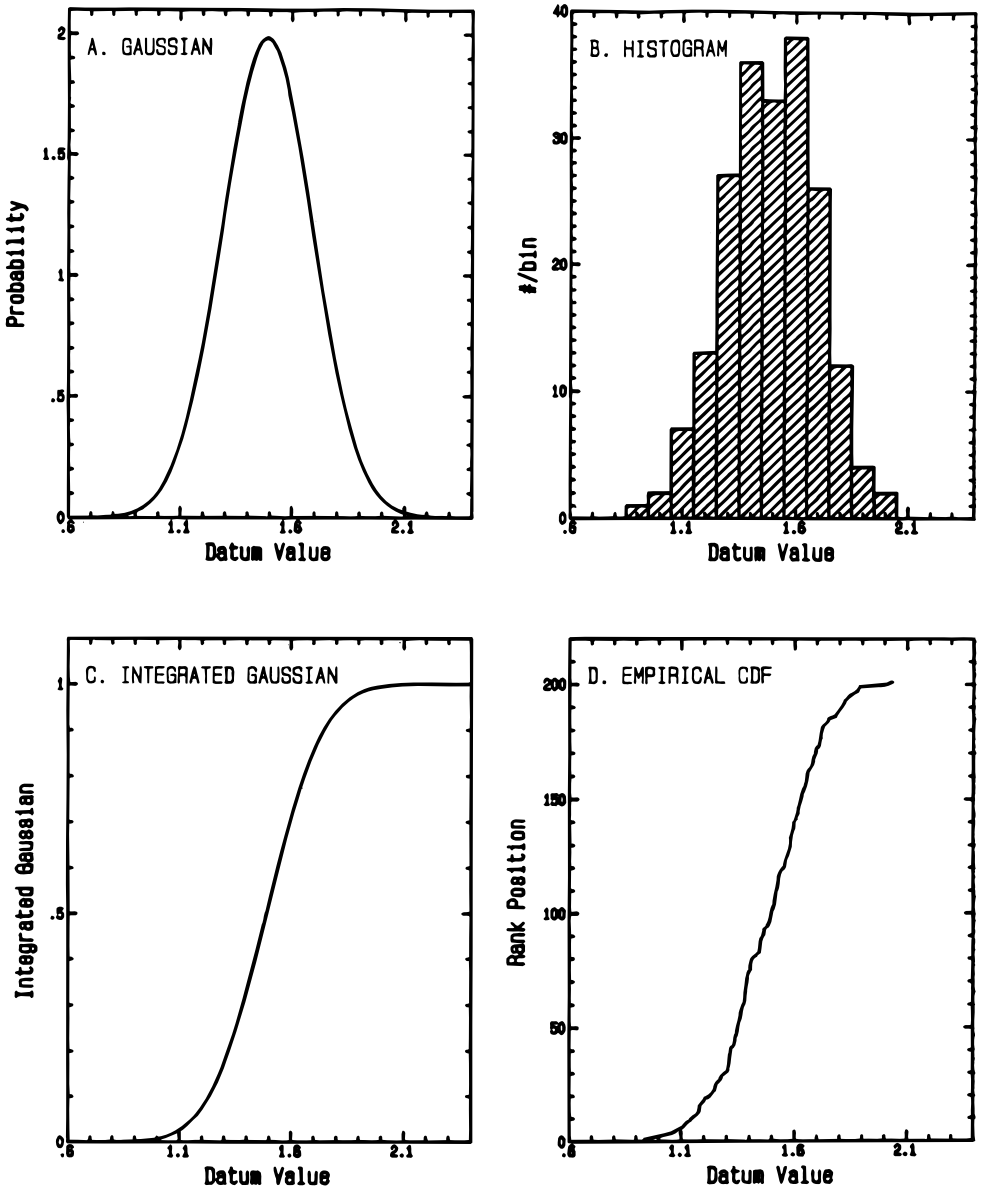

Figure 3. The top panel (A) shows the shape of a Gaussian (normal) distribution with a mean of 1.489 and a standard deviation of 0.201. This is compared in panel B to the histogram (bin size 0.1 units) of a 201 point random sample from such a distribution. Panel C shows the integrated Gaussian probability (integral = 1.0) for the function of panel A. This is compared to panel D, the empirical cumulative distribution function (CDF) for the 201 point random sample of panel B.

there are obvious and consistent differences at both the low and high ends of the distribution. These non-random, consistent differences indicate that the distribution is more complex. Fits with multiple Gaussian components using the DANFIP procedure give dramatic improvements in the reduced Chi-square up to four components. The four component fit is shown in Figure 5 (panel B) along with the magnified residuals (panel A) and the single Gaussian fit from Figure 4 (long dashed line, panel B).

The combined *B*-factor data of each structure has been analyzed using these tools. Analysis consists both of qualitative examination of the eCDF in graphical presentation (Figure 6) and quantitative examination by fitting it with a distribution model (Tables 4 and 5). All of the fits show a clear improvement in reduced chi-square with multiple components up to six (Table 4).

The distributions of Figure 6 in all cases are obviously not single Gaussians. Even the most symmetric looking of these curves (the 5LYT data set analyzed in Figure 4 and 5) is best fit by a 4 component model (Table 5). The quality of these fits is shown in the plots of the best (Figure 5) and worst (Figure 7). The 4LYTB fit (Figure 7) follows the general trends of the curve very well, but the residuals oscillate at the lower limit and in the central part of the plot. This indicates either that six components are not entirely sufficient or that Gaussian functions are not exactly appropriate in this case. However, it is clear that any additional components would be minor contributions. This can be seen in the magnitude of the residuals inferred from the accuracy of the fit line in following the data (main panel), i.e., the differences between the fit and the data are hardly discernable without magnification.

Another indication of how well these models fit the data is the comparison of the global means calculated from the weighted average of the mean values of the fit components (Table 1, last column) and the arithmetic means of the data (Table 1, third column).

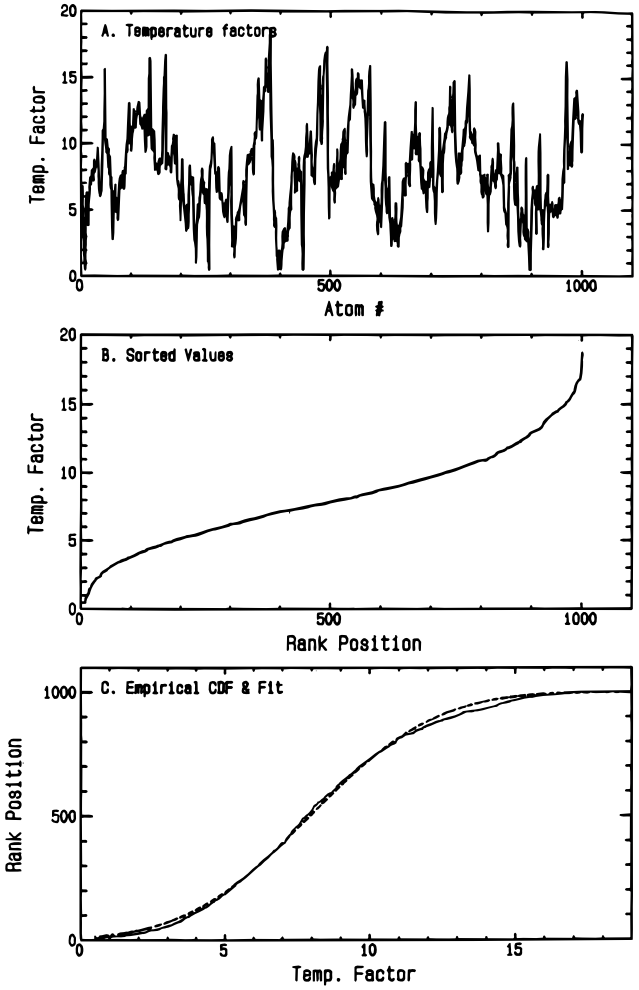

Figure 4. Generation of the empirical CDF illustrated with the 100 K tetragonal (5LYT) *B*-factors: panel A, the values plotted versus atom number; B, the sorted values, smallest value to the left, largest to the right; C. The empirical CDF generated from panel B by swapping the *X*- and *Y*-axes (solid line) and the best fit of a single scaled (to the number of points) integrated Gaussian probability (dashed line).

The eCDF is not a familiar plot to most experimentalist who are more familiar with histogram plots as shown in Figure 3. A histogram and an eCDF are related by differentiation and smoothing. Figures 5C and 7C show the histograms, derivative of the eCDF (short dashed line) and derivative of the best fit line (smooth, solid line) from Figures 5B and 7B, respectively. In Figure 5C, the single Gaussian that best fits this data set is also shown (long dashed line). In these plots the curves are all scaled to unit area to represent probability functions. The derivative of the eCDF is very noisy as expected since it does not have the smoothing of the histogram (from the width of the bins). However, it is clear from the Figures that the three curves (histogram, derivative of eCDF and derivative of fit) represent the same information. The approach of presenting the results of DANFIP analysis in terms of the derivative of the fit curve then gives us a more familiar form for viewing the distribution functions. In addition, the components of the fit can be more easily described by their individual bell shaped distributions scaled to fraction of total (Figure 8, panels B and C).

Table 4. Reduced χ^2 for Multicomponent DANFIP

structure file	temp (K)	reduced χ^2 for fits: 1 comp.	2 comp.	3 comp.	4 comp.	5 comp.	6 comp.
3LYTA	100[a]	933	38	20	17	11	7
3LYTB	100	1935	106	30	20	20	16
5LYT	100	148	15	15	6	6	6
4LYTA	298	1315	88	17	14	10	10
4LYTB	298	1148	122	46	27	8	7
6LYT	298	1758	211	23	12	7	7

[a] This data set was truncated at the lower end (50 identical *B*-factor values of 0.5 Å^2 and no values less than 0.5 Å^2). These values were removed for DANFIP analysis, and fits were obtained with the sample size (1001) retained (see ref 15 for more information on processing truncated data sets).

The fit models in the form of the derivatives of the fit function for all of the *B*-factor eCDFs are shown in Figure 8 (solid lines). The individual components shown (dashed lines) are tabulated in Table 5. In each case the area under the curve for a given component is the fractional contribution of that component (see Table 5).

DISCUSSION

The variation of *B*-factors in these data sets fit well to a model where an overall distribution of random values is sampled from a multicomponent function. The parent distribution function in each case can be effectively modeled as a combination of four to six normal or Gaussian distributions.

The obvious question is whether the individual components of these complex distributions correlate with particular groups or groupings of atoms. However, since the individual components overlap (Figure 8), most atoms can not be unambiguously assigned to one distribution alone. Even so, a course grained assignment can be made based on the

Table 5. Fits of Temperature Factor Data for All Atoms of the Hen Egg White Lysozyme Structures of Young *et al.*[14] [a]

file	temp (K)	dark blue component: mean	(SD)	%	light blue component: mean	(SD)	%	green component: mean	(SD)	%	yellow component: mean	(SD)	%	orange component: mean	(SD)	%	red component: mean	(SD)	%
3LYTA	100[b]	1.622	(1.601)	21.3	4.594	(1.164)	26.5	7.075	(0.597)	8.4	11.471	(0.205)	1.9	11.689	(3.044)	40.9	19.643	(0.560)	1.0
3LYTB	100	2.086	(1.271)	9.8	3.746	(0.652)	5.5	6.227	(2.210)	56.9	10.614	(0.495)	3.7	13.086	(2.276)	13.8	20.700	(1.482)	10.3
5LYT	100							6.107	(2.424)	57.5	7.538	(0.318)	2.9	9.858	(1.836)	29.0	14.277	(1.633)	10.6
4LYTA	298	8.989	(0.684)	3.0	14.949	(2.338)	54.1				20.184	(1.296)	15.0	23.540	(0.903)	4.9	26.483	(5.646)	23.0
4LYTB	298	7.617	(1.401)	8.3	12.689	(1.471)	39.8	15.902	(0.830)	13.9	19.172	(1.903)	24.1	24.226	(0.880)	2.5	29.038	(2.056)	11.4
6LYT	298	8.581	(1.024)	10.9	10.441	(0.685)	15.8	12.545	(0.670)	10.1	15.714	(3.489)	51.3				27.114	(3.308)	11.9

[a] Best fit is given based on reduced χ^2 values from Table 4. For each subpopulation shown in Figure 8, the mean (Å^2 units), standard deviation (SD), and the percentage contribution to the whole are given. In the case of four and five components, they have been aligned in the columns with the corresponding mean value component of the six component fits of that temperature. The color code of Figure 9 is used to lable the components. [b] This data set was truncated at the lower end (50 value = 0.5 Å^2, no values less than 0.5 Å^2). For DANFIP analysis these values were discarded, but the sample size (1001) was retained.[15]

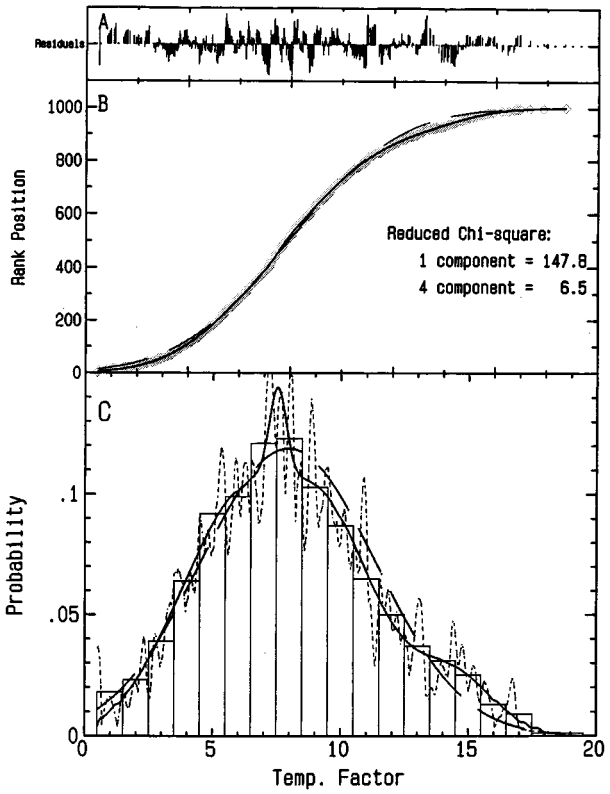

Figure 5. Improved fit to the 100 K empirical CDF of the 5LYT *B*-factors of Figure 4, panel C. The empirical CDF (shaded diamond symbols) and the fit (solid line) of a four population model (see Table 5 for values) are shown in panel B along with the single Gaussian fit of Figure 4, panel C (long dashed line). Panel A shows a considerably magnified plot of the residuals (differences) between the best fit and the eCDF of panel B. Panel C shows the scaled (to unit area) distribution functions in the form of the histogram (bar graph), the derivative of the eCDF (short dashes), the derivative of the best fit of a four population model (solid curve), and the derivative of the best single Gaussian model (long dashes).

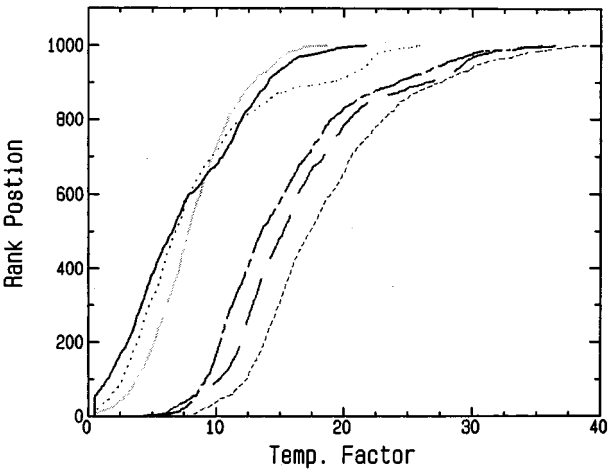

Figure 6. Empirical CDFs for the *B*-factors of all atoms in the hen egg white lysozyme structures. For the 100 K structures, the line types are as follows: solid line, monoclinic A structure (3LYTA); dotted line, the other 100 K monoclinic structure (3LYTB); long-shaded dashes, the 100 K tetragonal structure (5LYT). For the 298 K structures the line types are as follows: short dashes, monoclinic A structure (4LYTA); long dark dashes, other monoclinic structure (4LYTB); short long dashes, 298 K tetragonal structure (6LYT).

minima and inflection points of the overall distributions (Figure 8). These assignments are shown graphically in Figure 9 by color coding (specified also in Table 5). The

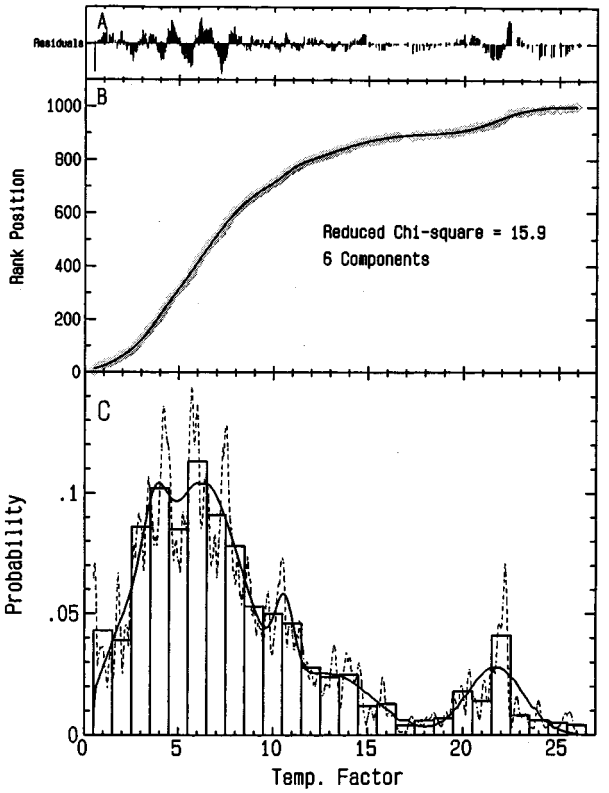

Figure 7. Six component DANFIP analysis of the second 298 K monoclinic structure (4LYTB). The empirical CDF (shaded diamond symbols) and the fit (solid line) of the six population model (see Table 5 for values) are shown in panel B. Panel A shows a magnified plot of the residuals (differences) between the best fit and the eCDF of panel B. Panel C shows the scaled (to unit area) distribution functions in the form of the histogram (bar graph), the derivative of the eCDF (short dashes), and the derivative of the best fit of a six population model (solid curve).

color code shows the components from that with the largest mean value to that with the smallest (red to dark blue respectively) truncated so that they do not overlap. However, with the 3LYTA model the narrow component centered on the broadest component (orange) is colored yellow and orange was assigned to the corresponding portions of the larger component on either side. It is obvious that there is not atomic or residue level correlation between the components from one data set to another. An obvious example is the tryptophane (TRP62) exposed in these views in the upper right part of the structure. This residue is variously assigned to all but the lowest *B*-factor component. Even in the structure pairs from the same crystal, its assignment varies. However, as is expected the more mobile exterior atoms are assigned to the higher mean value subpopulations and the interior atoms to the lower mean value subpopulations. In addition, in the orientation shown, the subpopulation with the highest mean *B*-factor is for the most part localized at the "north and south polar regions".

The atoms assigned roughly by their sub-population in this way tend to be group in a given structure. However, these groups do not correlate very well between structures, even between the two structures from the same crystal and same temperature. Indeed, with these assignments based roughly on the component subpopulations, the linear correlation coefficients at the atomic level are no better for the comparison between structures from the same crystal than in the comparison of the raw values (Table 3). In all cases

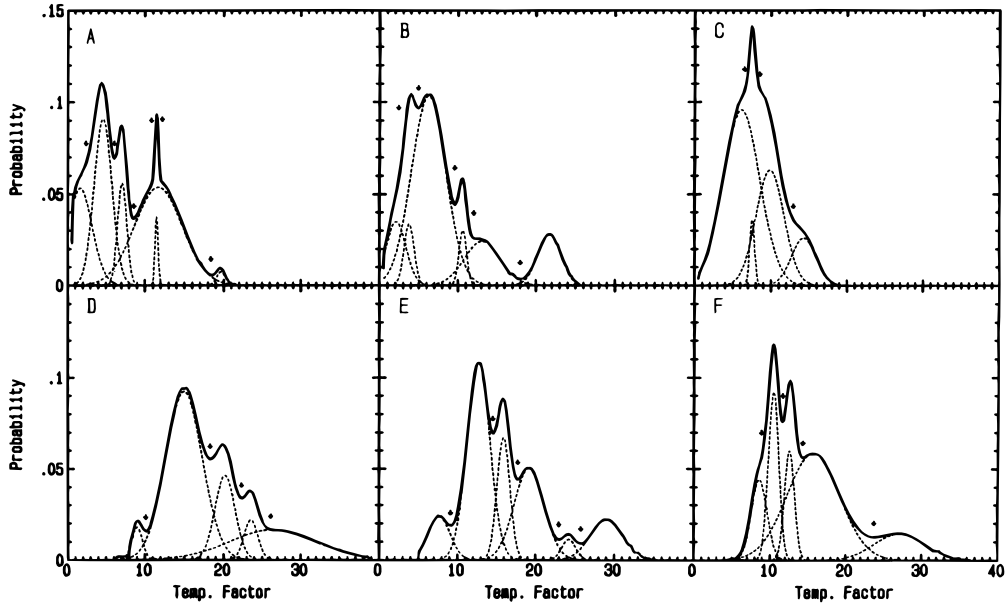

Figure 8. The first derivative of the best fit eCDF constructed from the components listed in Table 5 (solid lines) are shown scaled to unit area. The components describing the Gaussian subpopulations are given scaled to their corresponding contributions to the total (percents in Table 5): panel A, 3LYTA; panel B, 3LYTB; panel C-5LYT; panel D, 4LYTA; panel E, 4LYTB; panel F, 6LYT. The marking points (down arrows) indicate the locations of the breaks between color assignments for Figure 9.

except for two (5LYT versus 4LYTA and 4LYTB), the linear correlation coefficients are significantly lower when the grouped values of Figures 8 and 9 are assigned the same integer "*B*-factor" and then compared as in Table 3. In these two cases, the correlation coefficients are essentially unchanged (5LYT versus 4LYTA gave $r = 0.562$ and 5LYT versus 4LYTB gave $r = 0.519$). Other examples are, for the 100 K monoclinic pair (3LYTA vs 3LYTB), the linear correlation coefficient is 0.406 (compared to 0.565 from Table 3). For the 298 K pair (4LYTA vs 4LYTB), the value is 0.737 (compared to 0.807 from Table 3). To some extent the lowered value of correlation coefficient in these cases is explained by the very low resolution of the integer scale versus the continuous value scale of the actual data and by the overlap of the subpopulations which makes the classification of them as in Figures 8 and 9 imprecise.

The small, sharp component (yellow colored) in the 100 K structures, seems to be associated with terminal atoms of exterior side chains distributed more uniformly over the entire surface (except the poles). However, since this component overlaps large underlying components which are more locally distributed and about half or more of the values assigned to this yellow color must come from these other subpopulations, this conclusion is tentative.

Unlike the obvious global difference between the positions of the distributions on the *B*-factor scale seen in Figure 6, there is no clear temperature difference in the assignments of atoms to the various subpopulations. Remembering that the low value populations represented by light blue and dark blue in the lower panels (Figure 9) correlate in absolute value with the red and orange populations in the upper panel (Figure 9), the general effect of temperature seems to be a displacement of the entire distribution toward higher values with no dramatic change in assignments to the subpopulations. In other words, temperature increase in this case does not seem to change the flexibility or dynamic variation of atoms in the structure in any differential way. Other recent studies of temperature variation of structure and dynamics using *B*-factors an indicator[5,11] together offer equally general and ambiguous information. In the Tilton *et al*. study of ribonuclease A structures,[5] a roughly linear dependence of *B*-factors on experimental temperature is seen. However, in their data sets, like that of 3LYTA above, a significant part of the data are truncated to a lower limiting *B*-factor of 0.5 Å^2. This skews their results (Wampler, unpublished). In the Kurinov and Harrison study,[11] the arithmetic mean of the *B*-factors of lysozyme structures taken at six temperatures from 95 to 295 K show no change with temperature (the correlation coefficient of average temperature factor versus temperature = 0.222, Wampler, unpublished). In all of these cases, including the data reported on herein, there are experimental differences between the various data collections at different temperatures-- different mounting and preparation of crystals, different X-ray exposure times, even instrumentation and data processing differences. As a result it is difficult to separate the effects of temperature from these other factors. For example, with the data reported on here, all of the 100 K data sets were taken with crystals mounted on a glass fiber after removal of excess mother liquor. The 298 K data were taken from crystals in the mother liquor contained in capillary tubes. Room temperature radiation decay was noted for both crystal forms at 298 K but was only corrected for in the case of the monoclinic structures. Similarly, absorption corrections were differentially applied and different instrumentation was used for the 298 K tetragonal data collection than for the others. In the other studies mentioned above data collection times varied from 1 to 14 h with differing instrumentation and crystals mounted in both ways in each case. The tools described in this paper are now being applied to the analysis of all of these data sets in an effort to distinguish between temperature and other effects reflected in their *B*-factors.

As with the temperature effect, the differences between the two structures in each of the monoclinic crystals and the corresponding tetragonal structure show no clear pattern of variation in *B*-factor. By the measure of Stroud and Fauman[8]

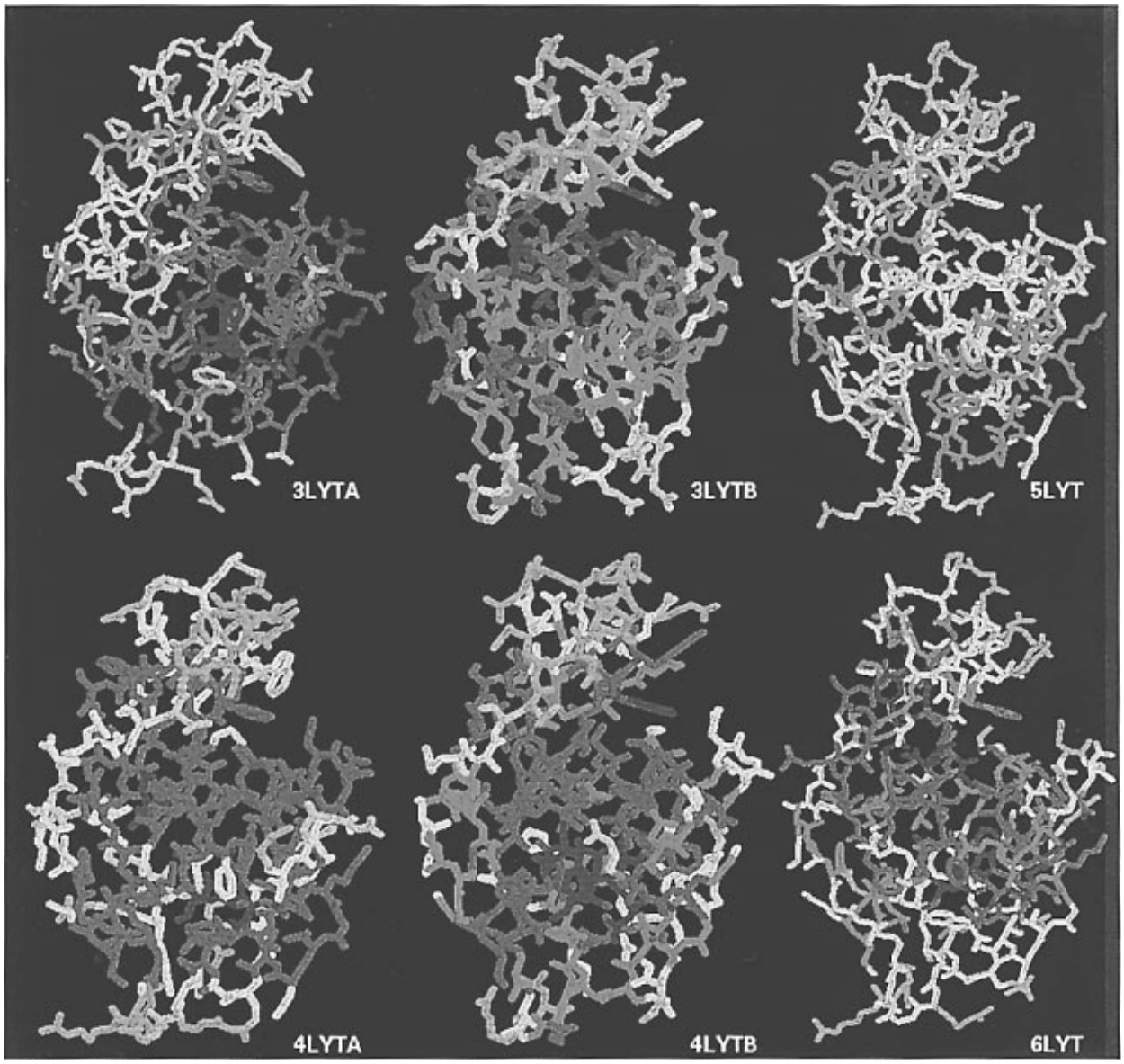

Figure 9. Structures colored according to the dividing points marked on the plots of Figure 8: red, highest mean value component; orange, next highest mean value component; yellow, third highest mean value component (with the 3LYTA structure this corresponds to the small, sharp component centered on the larger one). Green, light blue, and dark blue colos are assigned in order to each lower mean value component. In each case the lowest mean value component is dark blue except in the case of 5LYT where only four components are required.

all of these structures are well defined (atoms/reflection ∼ 0.2) predicting an RMS difference of about 0.2 Å at a low *B*-factor value of 0.5 Å^2 and a high of ∼0.7 Å at *B*-factor = 40 Å^2. Yet the correlations between differences in structural position and *B*-factors is poor. For a given atom in either of the two pairs of structures from the monoclinic crystals, the *B*-factor may vary in assignment between any of four to six widely separated subpopulations. The overall correlation between corresponding structures at the two temperatures in the monoclinic crystals is better than the correlation between the structures within the same crystal at the same temperature (Table 3). Indeed the best correlation is between 3LYTB and 4LYTB, suggesting that temperature effects on *B*-factors may be more correlated than those of other variables. A broader examination of the effect of temperature on *B*-factors is now underway. The Maxwellian distribution model for structural differences[13] may also be used with the DANFIP procedure (work in progress) to understand and correlate the variation between like structures and between structure variations and *B*-factors. Clearly, the statistics of such comparison should not be analyzed based on simple, single component distributions.

As mentioned in the introduction, there are few direct measures of atomically resolved motion on the fast time scale. *B*-factors are often used in comparison with computational measures of such motion from molecular dynamics simulations.[2,21] Indeed, one of the criticisms of such calculations is their failure to give very good correlations between experimental and calculated *B*-factors.[21] It should be clear from the data and analysis presented here that a failure of computationally evaluated mean square displacement to correlate well with experimental *B*-factors from crystallography may not reflect on the quality of the simulation at all. The examples cited by Hunenberger *et al*.[21] comparing calculated *B*-factors to experimental values with correlation coefficients from 0.45 to 0.85 are well within the range of the comparisons of experimental *B*-factors between similar structures (Table 3). Thus, the calculated

values correlate with experimental ones at a comparable level making assessment of the quality of the simulation results difficult based on this criterion. This conclusion is supported by the poor correlation between experimental *B*-factors and temperature in other crystallography studies mentioned above and the extremely wide range of *B*-factor data for homologous structures determined under comparable experimental conditions (Wampler, unpublished). Similar concern might also be expressed for comparisons between experimental measures. For example, Constantine *et al.*[22] found poor correlation between crystallographic *B*-factors and NMR measures of flexibility. These issues are being actively pursued in this laboratory.

ACKNOWLEDGMENT

This work was funded in part by grant GM50736 from the National Institutes of Health. The author acknowledges the helpful advice and discussions of Dr. Bi-Cheng Wang (this department). Thanks also to Dr. Wang and Dr. Cory Momany (also this department) for critical reading of the manuscript.

REFERENCES AND NOTES

(1) Sneddon, S. F.; Brooks, III, C. L. Protein motions: structural and functional aspects. In *Molecular Structures in Biology*; Diamond, R., Koetzle, T. F., Prout, K., Richardson, J. S., Eds.; Oxfrod University Press: Oxford, 1993; pp 114−163.

(2) van Gunsteren, W. F.; Hunenberger, P. H.; Mark, A. E.; Smith, P. E.; Tironi, I. G. Computer Simulation of Protein Motion. *Computer Phys. Comm*. **1995**, *91*, 305−319.

(3) LeMaster, D. M.; Kushlan, D. M. Dynamical Mapping of E. coli Thioredoxin via 13C NMR Relaxation Analysis. *J. Am. Chem. Soc*. **1996**, *118*, 9255−9264.

(4) Smith, J. C.; Kneller, G. R. Combination of Neutron-scattering and Molecular-dynamics to Determine Internal Motions in Biomolecules. *Molecular Simulation* **1993**, *10*, 363−375.

(5) Tilton, R. F.; Dewan, J. C.; Petsko, G. A. Effects of Temperature on Protein Structure and Dynamics. *Biochemistry* **1992**, *31*, 2469−2481.

(6) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. Protein Data Bank. In *Crystallographic Databases - Information Content, Software Systems, Scientific Applications*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.;. Data Commission of the International Union of Crystallography: Bonn, 1987; pp 107−132.

(7) Drenth, J. *Principles of Protein X-ray Crystallography*; Springer-Verlag: New York, 1994.

(8) Stroud, R. M.; Fauman, E. B. Significance of Structural Changes in Proteins: Expected Errors in Refined Protein Structures. *Protein Science* **1995**, *4*, 2392−2404.

(9) Petsko, G. A.; Ringe, D. Fluctuations in Protein Structure from X-ray Diffraction. *Annu. Rev. Biophys. Bioeng*. **1984**, *13*, 331−371.

(10) Frauenfelder, H.; Hartmann, H.; Karplus, M.; Kuntz, I. D., Jr.; Kuriyan, J.; Parak, F.; Petsko, G. A.; Ringe, D.; Tilton, R. F., Jr.; Connolly, M. L.; Max, N. Thermal Expansion of a Protein. *Biochemistry* **1987**, *26*, 254−261.

(11) Kurinov, I. V.; Harrison, R. W. The Influence of Temperature on Lysozyme Crystals. Structure and Dynamics of Protein and Water. *Acta Crystallogr*. **1995**, *D51*, 98−109.

(12) Chothia, C.; Lesk, A. M. The Relation Between the Divergence of Sequence and Structure in Proteins. *EMBO Jour*. **1986**, *5*, 823−826.

(13) Chambers, J. L.; Stroud, R. M. The Accuracy of Refined Protein Structures: Comparision of Two Independently Refined Models of Bovine Trypsin. *Acta Crystallogr*. **1979**, *B35*, 1861−1874.

(14) Young, A. C. M.; Dewan, J. C.; Nave, C.; Tilton, R. F. Comparison of Radiation-Induced Decay and Structure Refinement from X-ray Data Collected from Lysozyme Crystals at Low and Ambient Temperatures. *J. Appl. Cryst*. **1993**, *26*, 309−319.

(15) Wampler, J. E. Analysis of the Probability Distribution of Small Random Samples by Nonlinear Fitting of Integrated Probabilities. *Analytical Biochemistry* **1990**, *186*, 209−218.

(16) Bradley, J. V. *Distribution-Free Statistical Tests*; Prentice-Hall: Englewood Cliffs, NJ, 1968; pp 15−44.

(17) Cooper, R. A.; Weekes, A. J. *Data, Models and Statistical Analysis*; Barnes and Noble Books: Totowa, NJ, 1983.

(18) Furukawa, R.; Wampler, J. E.; Fechheimer, M. Cytoplasmic pH of *Dictyostelium discoideum* Amebae during Early Development. *J. Cell Biol*. **1990**, *110*, 1947−1954.

(19) Wampler, J. E. SPECOS SA, a Computer Program for Managing, Graphing and Manipulating Laboratory Data. *Analytical Inst*. **1990**, *19*, 203−230.

(20) Sayle, R.; Milnerwhite, E. J. RASMOL-Biomolecular Graphics for All. *Trends Biochem. Sci*. **1995**, *20*, 374−376.

(21) Hunenberger, P. H.; Mark, A. E.; van Gunsteren, W. F. Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J. Mol. Biol*. **1995**, *252*, 482−503.

(22) Constantine, K. L.; Friedrichs, M. S.; Goldfarb, V.; Jeffrey, P. D.; Sheriff, S.; Mueller, L. Characterization of the Backbone Dynamics of an Antidigoxin Anitbody-VL Domain by Inverse Detected H-1-N-15 NMR Comparisons with X-ray Data for the FAB. *Proteins* **1993**, *15*, 290−311.

CI9702252