Proc. Indian natn Sci Acad.(PINSA) **B66** Nos 4 & 5 pp239-264 (2000)

Protein Dynamics by X-ray Diffraction: An Analysis of Atomic Displacement Parameters (ADP), the B-values

S PARTHASARATHY and M R N MURTHY*
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012

(Received on 2 February 2000; Accepted after revision on 17 July 2000)

The temperature factors or atomic displacement parameters (ADPs or B-values) obtained from high-resolution X-ray refinement of proteins represent the flexibility of the polypeptide. A critical analysis of these ADPs, therefore, provides insights in to the relation between protein dynamics and its sequence and structure. Further, comparative analysis of ADPs in homologous proteins could provide information on constraints on protein evolution imposed by pressures to preserve the dynamics.

ADPs show large variations from one structure to another. It is shown that the B-values, when expressed in units of standard deviation about their mean value (B'-factors) at the Cα atoms, have a characteristic frequency distribution. All analyses described here have been performed using these modified B'-factors. The frequency distribution of B'-factors in each protein could be modeled as a sum of two Gaussian functions. The parameters describing the functions were found to be characteristic properties of proteins. The frequency distribution for a given amino acid over all the proteins also showed a similar pattern. The ratio of the area under the two functions could be used as a parameter representing the amino acid propensity to occur in rigid / flexible regions of the polypeptide.

Analysis of the thermal parameters of proteins from thermophilic and mesophilic organisms revealed that Ser and Thr have lesser flexibilities in the former. In regions of the polypeptide with high B-values, compositions of Glu and Lys were higher and those of Ser and Thr were lower in thermophiles when compared to mesophiles.

Comparative analysis of the ADPs in homologous proteins indicated that the flexible and rigid regions in the three-dimensional fold remain largely conserved during the course of evolution. The variation in the flexibility at a given position was only weekly correlated to the variation of the amino acid sequence at the corresponding position. These results illustrate that the relationship between sequence and dynamics has degeneracy similar to that of sequence and three-dimensional structure.

Statistical analysis of the relationship between the flexibility of the protein molecule and its conformation revealed that the ADPs of side chain atoms are lower for energetically favorable rotamers. Thermal parameters also varied with non-planar distortions of the peptide as represented by the ω angle. The conformations with ω larger than the ideal trans geometry (180°-190°) were more flexible when compared to those with ω < 180° (170°-180°). The average ADPs of peptide units were found to depend weekly on the Ramachandran angles at corresponding Cα atoms.

B'-factor frequency distributions and correlation coefficients (CCs) between the mean B-values of main chain and side chain atoms could be used to assess the reliability of B-values. The distribution of CCs showed dependence on the package used for refinement (X-PLOR, PROLSQ or TNT). Further differences were also discernible in proteins refined using the same package. The importance of these observations for validating protein refinement is discussed.

Key Words: Proteins, Atomic displacement parameters, Temperature factors, Statistics, Gaussian functions, Accessibility, X-ray structures, Protein stability, Protein dynamics, Thermophiles, Mesophiles, Protein refinement, Validation, Conformation, Evolution

Author for Correspondence: **Tel**: 080-3091612, **Fax**:080-3341683, **E-mail**: mrn@mbu.iisc.ernet.in

Introduction

With rapid advances in experimental techniques and theoretical methods, X-ray structure analysis of proteins has changed from providing static images of biological macromolecules to a powerful technique for the interpretation of function and stability in terms of dynamics (*for reviews see* Petsko & Ringe 1984, Ringe & Petsko 1986, Rejto & Freer 1996). All functions carried out by proteins inevitably involve atomic movements, from small shifts of the order of fraction of an Å to larger changes of the order of several Angstroms. The time scales of these motions range from femto seconds to a few milli seconds. Motions in proteins encompass; (i) small atomic displacements required for the binding of substrates or co-factors, (ii) motion of loops resulting in the creation of an environment suitable for chemical function, (iii) relative movement of domains caused by hinge bending etc. These motions are essential for a variety of functions such as "induced-fit" of enzyme and substrate (hexokinase), stabilization of transition states (serine proteinses), hinge motions of domains (citrate synthase), cooperativity in function resulting from movements of whole subunits (haemoglobins), metal ion binding (calcium binding to EF-hand proteins) and protein-protein association (virus particle assembly).

A variety of spectroscopic, diffraction and computational techniques have been used to understand the motions associated with protein molecules. Each technique is unique in terms of the nature and time scale of information it provides. This review deals mainly with our published work on atomic motion as deduced from X-ray diffraction studies on single crystals of proteins. No attempt will be made to cover the extensive literature available on the application of other techniques.

Flexibility in Protein Crystals

In protein crystals, individual molecules are packed with only a small number of inter molecular interactions. The solvent used for crystallization fills about 40-60% of the crystal space. Because of weak intermolecular interactions and high solvent content, most of the motions available for protein molecule in solution are also available in the crystalline state. Thus, protein crystals can be considered as concentrated solutions. The evidence for flexibility in protein crystals has come from several sources, viz. (i) comparable activity of several enzymes in solution and crystalline state, (ii) reversible binding of oxygen by haemoglobin crystals and (iii) H-exchange studies on lysozyme crystals. It is now well established that the motions of atoms within a crystalline lattice as deduced by X-ray techniques agree with values obtained by NMR methods for the same protein in solution.

X-ray Diffraction and Dynamic Information

Early X-ray diffraction studies were confined to the elucidation of a "rigid structure" for the poly peptide fold. With the advent of sophisticated techniques for the accurate measurement of intensity data and computational power for the refinement of protein structures, this "rigid" view of proteins has been replaced by that of molecules with characteristic dynamics.

The X-ray crystal structures are averages over the time of data collection and over conformational states the molecule might assume in different unit cells. Compared to other methods, X-ray diffraction is unmatched in its ability to provide spatial information. The flexibility of atoms in these experiments appears as distortions of experimental electron density. The extent of distortion depends on both the amplitude of atomic fluctuations and on the resolution of the available data. Distortions in electron densities obtained using high resolution (better than 2Å) data can be interpreted in terms of individual atomic displacements. Atomic fluctuations are quantified in terms of Atomic Displacement Parameters (ADP, B-values), described less precisely also as temperature factors or thermal factors.

Effect of Atomic Displacements on Bragg Intensities

Intensities of Bragg reflections from real crystals fall of more rapidly than the theoretical estimate based on electron density distributions of static atoms. This reduction in intensities is a result of static or dynamic features of crystalline state. Static effects arise if molecules assume alternate

conformations in different unit cells. The dynamic effects are primarily due to vibrations of atoms about their mean position. These disorders are more prominent in protein crystals than in crystals of small molecules because of week intermolecular interactions and high solvent content of protein crystals. The extent to which protein crystals diffract gets severely limited due to these disorders. Frequencies of atomic vibrations are of the order 10^{13} per second whereas those of X-rays are $2x10^{18}$ per second. Hence, X-ray scattering appears to be due to stationary atoms displaced randomly from their mean positions. This is equivalent to scattering by "smeared atoms" resulting from spatial spread of atomic electron density. In 1914, Debye showed that the scattering factor for "smeared" atoms could be expressed as,

$$f=f_0 exp(-B\ sin^2\theta/\lambda^2)$$

where f_0 is the atomic scattering factor of a static atom (with value at zero Bragg angle equal to the number electrons) and f is the scattering factor corrected for atomic motion. The exponential term describes the decrease in scattering amplitude at higher Bragg angles. The B-value appearing in the exponential term is called Atomic Displacement Parameter (ADP), also popularly referred to as the temperature or thermal factor.

Figure 1 shows the effect of increasing B-values on atomic scattering factor plotted against $\sin\theta/\lambda$. Because of this reduction, intensities reduce to noise level above some diffraction angle θ_{max}. From Bragg's equation, effective resolution of the structure determined is $d_{min} = \lambda / 2sin\theta_{max}$. Thus, for crystals with overall B-value of 40Å^2 or greater, it is not possible to get diffraction data better than 2Å resolution.

Models of B-Values of Protein Crystals

Estimation of Over all B-value of the Crystal

Wilson statistics provides the means of estimating the overall B-value. For a structure with randomly distributed atoms, Wilson (1949) showed that the mean intensity in shells of increasing resolution satisfy the equation

$$\langle I \rangle = k * \Sigma f_i^{\,2} * exp\ (-2Bsin^2\theta / \lambda^2)$$

where $\Sigma f_i^{\,2}$ represents the sum of the squared atomic scattering factors over all the atoms of the unit cell, B is the mean atomic displacement parameter and $1/k$ is the scale factor required to place the observed intensities on an absolute scale. A plot of $ln\ (\langle I \rangle / \Sigma f_i^{\,2})$ against $sin^2\theta / \lambda^2$ provides estimates of k and B in terms of the intercept and slope of the least squares line. Although atoms of protein structures are not randomly distributed, Wilson plot (figure 2) has been found to provide a reasonable estimate of the overall atomic displacement parameter. Corrections for both the relative scale and atomic displacement parameters should be applied to merge and scale intensities of Bragg reflections from two independent data collection steps.

Harmonic Oscillator Approximation

The extent to which the atoms in crystals undergo

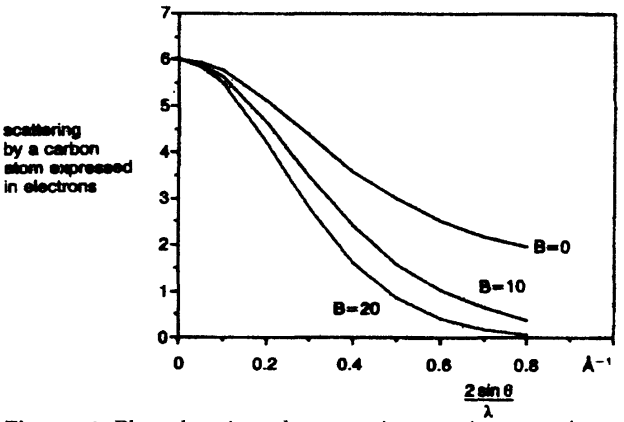

Figure 1 Plot showing decrease in atomic scattering factor of carbon atom with increasing B-values*

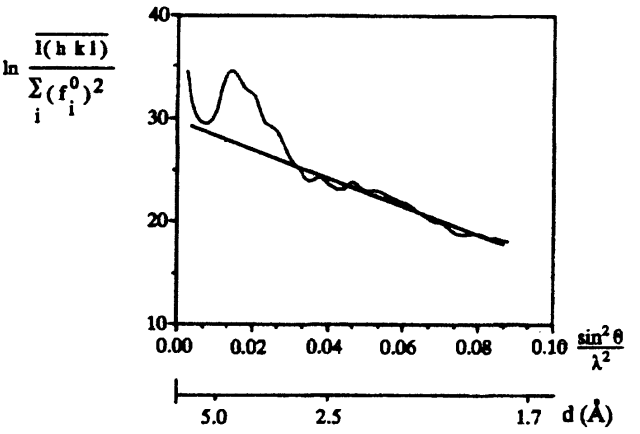

Figure 2 The Wilson plot for phospholipase A2 with data to 1.7Å resolution*

*Reproduced with permission from *Principles of Protein X-ray Crystallography* by Jan Drenth 1994 Springer-Verlag, New York, Inc

displacements from their equilibrium positions depends on the force field exerted by the surrounding microenvironment. The motions of the individual atoms, thus, are coupled to those of the neighboring atoms. The description of these correlated motions is complicated. A simple approximation is to assume that the atoms move in a harmonic potential. The time-averaged movement of an atom in a harmonic potential is described by a Gaussian probability density function. Movement of such harmonic oscillators can be approximated to isotropic or anisotropic displacements.

Isotropic Model

In this approximation, the expression for scattering power of an atom can be written as,

$$f = f_0 \exp(-2\pi^2 \langle \mathbf{u}^2 \rangle \mathbf{hTh})$$
$$= f_0 \exp(-8\pi^2 \langle \mathbf{u}^2 \rangle (\sin^2\theta / \lambda^2))$$
$$= f_0 \exp(-B (\sin^2\theta/\lambda^2)) \quad \rightarrow (2)$$

where, $\langle \mathbf{u}^2 \rangle$ is the average of mean square atomic displacements along the three axes, i.e.

$$\langle \mathbf{u}^2 \rangle = (u_x^2 + u_y^2 + u_z^2)/3$$

h is a reciprocal lattice vector, θ is the scattering angle and λ is the X-ray wavelength. In macromolecular crystallography, the atomic displacements are normally expressed as *B*-values instead of mean square displacements. The *B*-values are related to mean square displacement of atoms as

$$B = 8\pi^2 \langle \mathbf{u}^2 \rangle$$

Expression (2) describes an electron cloud uniformly smeared in all the directions and hence the displacement parameter derived from it is isotropic. In X-ray structure analysis, intensities constitute the observables for the refinement of both positional and displacement parameters. Refinement is severely limited by low observable to parameter ratio in protein structure determinations due to the large number of atoms in protein molecules. This ratio strongly depends on the diffraction data resolution. This limitation is overcome in practice by the use of constraints or restraints on the geometry of protein molecules. The description of individual isotropic ADP requires four parameters (three positional and one displacement) for each atom. It is meaningful to refine individual isotropic ADPs only when the available diffraction data extend to a resolution of 2.5Å or better. At lower resolutions of 3.0 Å or less, only the over all B-factor or grouped B-factors might be refined.

Anisotropic Model

The smearing of electron density is, in general, not equal in all the directions. Anisotropic displacements of atoms might be described by a three-dimensional Gaussian probability density function based on harmonic potentials. Under these conditions the **u** in expression 2 will be a 3x3 symmetric tensor. The structure factor equation becomes,

$$f = f_0 \exp(-2\pi^2 \mathbf{hTUh})$$
$$= f_0 \exp[2\pi^2 (h^2U_{11}(a^*)^2 + k^2U_{22}(b^*)^2 + l^2U_{33}(c^*)^2 + 2hkU_{12}(a^*)(b^*) + 2hlU_{13}(a^*)(c^*) + 2klU_{23}(a^*)(c^*))].$$

The six independent components of the tensor Uij represent the anisotropic displacements of the atom. The three-dimensional Gaussian function contoured at a given probability value is called the displacement ellipsoid. Switching the description of individual atomic *B*-values from isotropic to anisotropic displacements increases the number of parameters to be refined per atom form four to nine and hence drastically reduces the observable to parameter ratio. Therefore, Anisotropic *B*-value refinement is not usually possible for protein crystals. Recent advances, however, in data collection techniques and radiation sources, have led to the possibility of collecting near-atomic and atomic resolution data on a few protein crystals, thereby improving the observable to parameter ratio. This has allowed the refinement of anisotropic *B*-values.

Interpretation of B-Values in Protein Structures

The authors of the TNT refinement program have suggested different levels of ADP refinement, depending on the resolution of the data available. It is not worth refining individual *B*-values when the resolution of the data is worse than 3.0 Å. Refinement of B-values with restraints could be carried out when the data is between 2.5 to 3.0 Å. Unrestrained refinement will be possible only when the data extends beyond 2.0 Å. Whenever the data extends beyond 2.5 Å, it is generally possible to locate alternate conformations of the

structure by careful inspection of Fo-Fc difference Fourier maps. The features of electron density that are not explained by the refinement of only atomic positions in these maps could be accounted for by careful refinement of *B*-values. The parts of the model that do not fit the density well, generally, have higher *B*-values than the average value for the rest of the model. One can resort to refinement of occupancies in alternate conformational states to account for unexplained features remaining after the *B*-value refinement. Refinements of *B*-values and occupancies should be carried out with caution as these two are highly correlated. The *B*-values model the ADPs while the occupancies model the disorders that are observed over different unit cells. Recently, a number of protein structures have been determined at atomic resolution, i.e. better than 1.2 Å (Dauter et. al. 1997, Longhi et al. 1998). This has allowed full-matrix least squares refinement of protein structures as well as anisotropic modeling of *B*-values. Anisotropic refinement improves the agreement between the model and the data and leads to a drop of about 5% in R-factors. This allows interpretation of subtle movements of atoms and determination of protonation states at the active site and hence provide valuable insights on protein function (Longhi et. al. 1997), detection of alternative conformations of side chains and unusual hydrogen bonding patterns and distances (Kuhn et al. 1998). The unrestrained (both stereochemical and *B*-value) refinements of atomic resolution structures are also useful for deducing better parameters required for restrained refinement at lower resolution and molecular dynamics studies. Merritt (1999a & b) has analyzed the anisotropic ADPs of atomic resolution structures and derived anisotropic *B*-value constraints that could be used at low resolution. Possibility of using mixed anisotropic (for well-defined atoms) and isotropic *B*-values (for other atoms) have also been discussed (Dodson et al. 1996). Libeu et al. (1997a) used such a procedure in the refinement of a mutant of Pseudoazurin and provided structural explanation for the observed redox potentials. They have also suggested a displacement parameter weighted co-ordinate comparison for detection of significant structural differences between oxidation states of the same protein (Libeu et al. 1997b).

ADPs are critical parameters for understanding function and dynamics of proteins. Here, we review our work on the use of ADPs for understanding a variety of properties related to proteins, viz., (i) statistical analysis of *B*-values with a view of determining flexibility indices of amino acids and their use in predicting antigenic regions of a given sequence (Parthasarathy & Murthy 1997), (ii) comparison of *B*-values of thermophilic and mesophilic organisms to obtain insights into protein thermal stability (Parthasarathy & Murthy 2000a), (iii) examination of the B-values in homologous protein families with a view of estimating the degree of conservation of dynamics during the course of evolution (Lenin et al. 2000), (iv) relation between B-values and conformational states of protein structures (Parthasarathy & Murthy 2000b), (v) correlation between B-values of main chain and side chain atoms which was found to indicate the need for adequate *B*-value refinement strategies and finally (Parthasarathy & Murthy 1999), (vi) application of frequency distribution of *B*-values as a possible validation tool in protein refinement. The protein structures used in these studies were chosen from PDB's (Brenstein et al. 1977) Nov. 1996 release of representative list (Hobohm & Sander 1994). The maximum sequence similarity between any two of these structures is 25%. All these structures were determined at a resolution better than 2.0 Å and have R-factors less than 20%. Ninety five of these were selected for analysis.

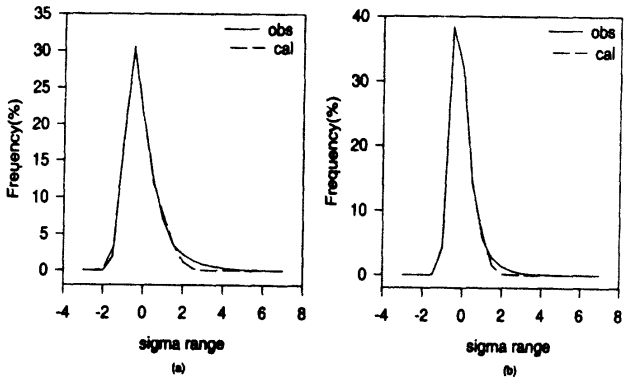

Figure 3 Frequency distribution expressed as percentages in bins of 0.5 (a) and 0.75 (b) units of B' over 95 high-resolution protein structures. Continuous lines correspond to observed frequencies and broken lines represent the sum of the two Gaussuan functions fitted to the observed distribution

Frequency Distribution of *B*-values in Proteins and of Individual Amino Acids

Mean Profile of B'-Factors Over all the Structures

For each selected protein, the mean of B at $C\alpha$ positions was calculated as

$$\langle B \rangle = \Sigma B_i / N,$$

where B_i is the B-value associated with $C\alpha$ of the i^{th} residue and N is the total number of residues in the protein. Similarly, the variance was computed as

$$\sigma^2(B) = \Sigma (B-\langle B \rangle)^2 / N;$$

Modified B (B'-factor) representing normal variate was computed as

$$B' = (B-\langle B \rangle) / \sigma(B)$$

The frequency distribution was sampled in bins of 0.5 in B'. This frequency distribution was obtained for all amino acids in individual proteins and for individual amino acids over all the proteins. The observed frequency distribution was then fitted to the following sum of two Gaussian functions,

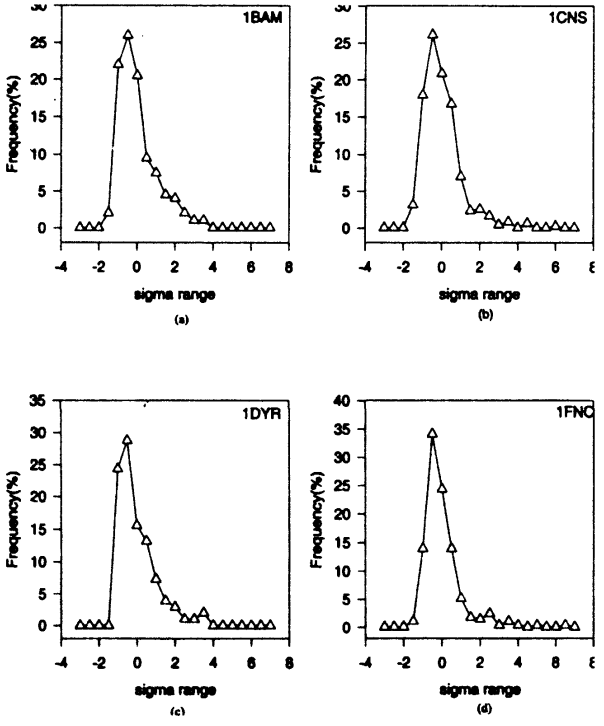

Figure 4 Frequency distribution expressed as percentages in bins of 0.5 units of B' for some representative high resolution protein structures: (a) Restriction endonuclease Bam H1 (1BAM), (b) Chitinase (1CNS), (c) Dihydrofolate reductase (1DYR) and (d) Ferredoxin (1FNC)

Table 1 *Amino acid composition corresponding to B' range – 0.75 to -0.25 and 0.25 to 0.75 in the 95 high resolution protein structures. Also listed is the over all composition in these proteins*

Amino acid	% composition in the range -0.75 to–0.25	0. 25 to 0.75	% composition over all
Ala	9.4	8.5	8.8
Cys	1.5	1.3	1.4
Asp	4.9	6.5	5.8
Glu	4.2	7.0	5.3
Phe	5.1	2.8	4.0
Gly	7.4	7.9	8.2
His	2.5	1.8	2.3
Ile	6.1	4.4	5.5
Lys	4.1	8.3	5.6
Leu	8.7	6.3	7.8
Met	2.5	1.3	2.0
Asn	4.4	5.5	4.9
Pro	3.6	5.8	4.4
Gln	3.3	4.0	3.7
Arg	4.4	4.1	4.1
Ser	6.6	7.9	7.0
Thr	6.6	7.0	6.7
Val	8.0	5.9	7.0
Trp	1.9	1.0	1.6
Tyr	4.6	2.6	3.9

$$f = k1\, e^{-k2\,*(B'-B1)^2} + k3\, e^{-k4\,*(B'-B2)^2}$$

where, *k1, k2, k3, k4, B1* and *B2* are parameters describing the two Gaussian functions. These parameters were refined by a least-squares procedure

The observed and least square fitted frequency distributions for bin sizes of 0.5 and 0.75 are shown in figure 3. The plot corresponds to a total of 31,918 residues in the 95 selected proteins. The fitted curve agrees very well with the observed distribution. The average *k1* and *k3* values, representing the peak height of the Gaussians, are 25.12 and 8.53, respectively. Similarly average *k2* and *k4*, representing the width of the Gaussians, are 0.55 and 0.16, respectively. The mean positions (derived from *B1* and *B2* values) for the two

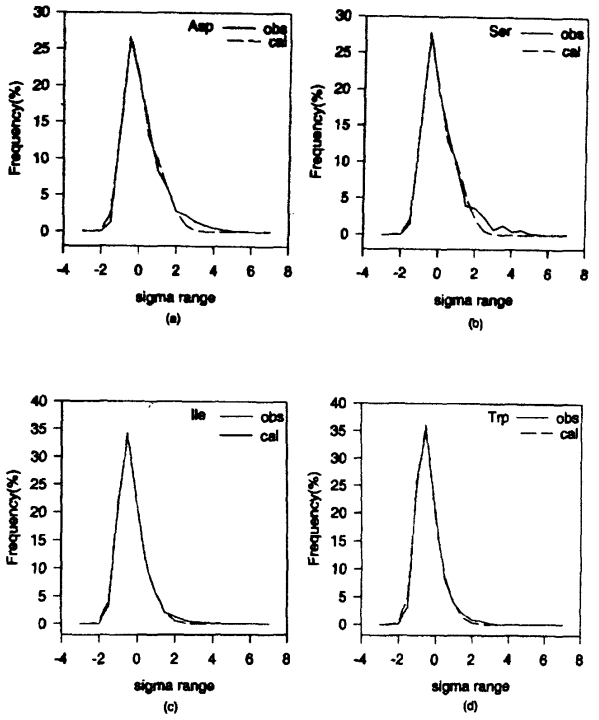

Figure 5 Frequency distribution expressed as percentages in bins of 0.5 units of B′ for representative amino acid residues: (a) Asp, to represent charged residues, (b) Ser, to represent small polar residues, (c) Ile, to represent aliphatic, hydrophobic residues and (d) Trp, to represent aromatic residues

Gaussians are, -0.55 and 0.39 for bin size of 0.5. These values are slightly different for bin size of 0.75 (-0.34 and 0.34). The results, presented hereafter, correspond to a bin size of 0.5.

The *B′*-factor frequency distribution is characteristic of the protein structures independent of their function and size as illustrated in figure 4 (a-d) for selected protein structures; Endonuclease (1BAM), chitinase (CNS), Dihydrofolate reductase (1DYR) and Ferridoxin (1FNC). In the 95 proteins used for this analysis, values of *B1* and *B2* varied within 10% for 93(98%) and 91(96%) and within 5% for 76 (79%) and 80 (84%) structures, respectively. Only for two structures, 2TGI and 4FGF, the constants were very different. This variation could be attributed to the special dynamic features of these proteins (see later).

Table 1 gives the amino acid composition over all the 95 proteins in the bins corresponding to the peaks of the two Gaussians, along with the overall amino acid composition in these proteins. It may be seen that the composition corresponding to the peak at lower B′-factor is rich in hydrophobic residues when compared to the average composition. Accordingly, the composition corresponding to the peak at larger B′ is richer in hydrophilic amino acids.

Frequency Distribution for Individual Amino Acid Residues

Figure 5 (a-d) shows the B′-factor frequency distribution for selected residues, considering all the 95 proteins. The fit of two Gaussian functions to the distribution of every residue type was found to be good. Lowest value of B1 was found for His (–0.69), and the highest value was found for Lys (-0.31). Similarly, the highest value found for B2 was 0.94 (Trp) and the lowest value was 0.60 (Lys). This observation suggests that protein structures are likely to exhibit characteristic *B′*-factor frequency distribution as far as they have some "universal" composition. Using the least-sequares fitted constants, the areas, *A1* and *A2*, under the 1st and 2nd Gaussian functions were calculated for individual amino acids as,

$A1 = k1 * (\pi / k2)^{1/2}$

$A2 = k3 * (\pi / k4)^{1/2}$

The fractional area under the second Gaussian for each amino acid, $p = A2 / (A1+A2)$ is likely to represent the probability or propensity that the given amino acid occurs with high *B′* value.

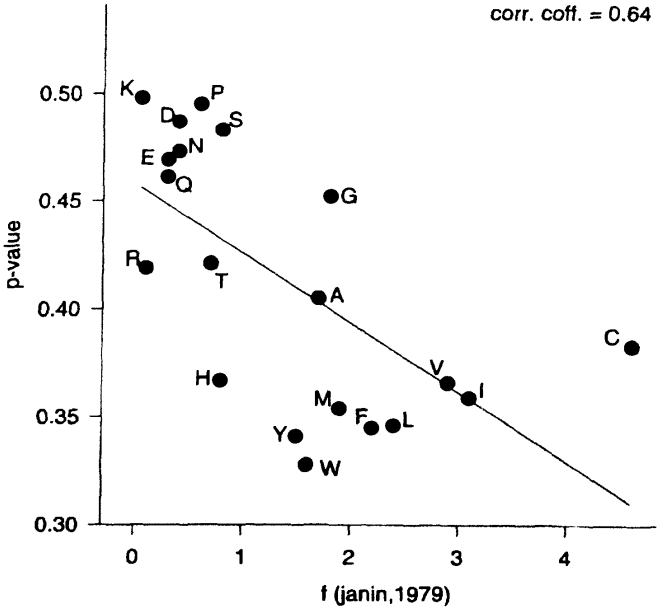

Figure 6 Plot showing correlation between *p* values and *f* values of Janin (1979)

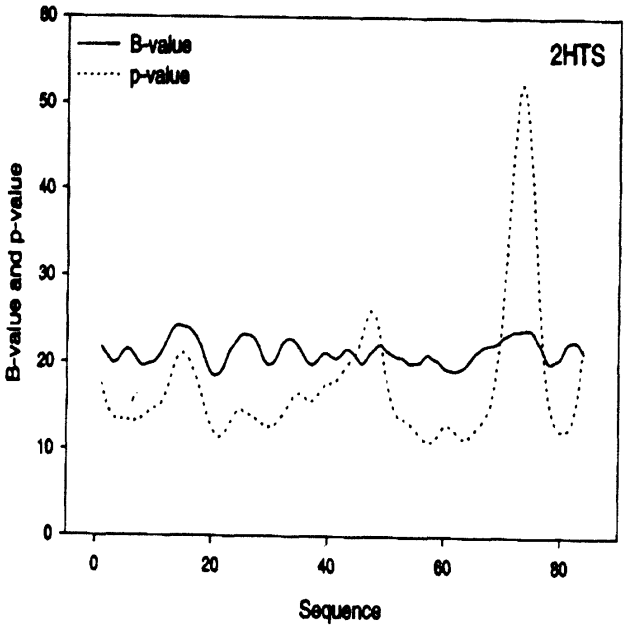

Figure 7 Profile of *p* and *B*-values (on arbitrary scales) averaged in windows of size 5 for DNA binding domain of heat shock transcription factor (2HTS)

Janin (1979) has derived a hydropathy scale based on fractional solvent accessibility of amino acids in proteins. Figure 6 represents the relationship between the fractional area under the 2nd Gaussian, (*p* values) and the fractional solvent accessibility parameter (*f*) of Janin (1979). The two parameters have a correlation coefficient of 0.64. Residues that show significantly lower *p* value than that anticipated from a linear relationship to Janin parameters are mainly Met and the aromatic residues Phe, Tyr, Trp and His. This suggests that these residues confer rigidity to the polypeptide segment. In contrast, the fraction of Gly, Pro and Thr residues that occur in regions of high *B*'-factors are more than their mean fractional exposure in proteins. Thus, the hydropathy index does not fully account for the probability of occurrence of the amino acids in high B-value segments.

Predicting Flexible Regions of Polypeptide Chains

The *p* values represent the tendency to occur in regions of high *B*'-factor. Therefore, a plot of *p*-value against the residue number in any protein of known sequence might reveal segments of its three dimensional structure that are likely to show large mobility. Figure 7 shows the plot of *p*-values against the sequence for the DNA binding domain of heat shock transcription factor (2HTS). The values shown are averages over a window of size 5. Observed *B*-values (averaged over a window size of 5) for the same protein are also shown in the figure. The agreement between the *p*-values and *B*-values is 57%. The mean correlation coefficient between the *p* values and the *B*-values over 94 proteins is 0.32 and is as high a value as reported in earlier investigations (Karplus & Shulz 1985, Regone et. al. 1981, Vihinen et al. 1994). Interestingly the membrane protein porin, 2POR, shows negative correlation. The profile of *p* values is therefore a useful tool in predicting flexible segments in proteins. These segments are also likely to be antigenically active.

Comparison of B-values in Thermophilic and Mesophilic Proteins

Inferences on Protein Thermal Stability

Increasing the thermal stability of proteins is one of the primary goals of protein engineering. Sequence comparisons between proteins from mesophilic and thermophilic origins have been performed in order to gain information on mutations that might lead to increased thermal stability. Merklet et al. (1981) correlated hydrophobic index and the Arg/(Lys+Arg) ratio to the optimum growth temperature of organisms. Arias and Argos (1989) suggested "traffic rules" to improve thermal stability. In their analysis, Lys→Arg and Ser→Ala appeared to be the top two amino acid substitutions (in helical segments) from mesophilic to thermophilic protein sequences. However, based on sequence comparison of dihyrofolate reductase and glutamate dehydrogenase sequences, Bohm and Jaenicke (1994) concluded that the proposed 'traffic rules' are not adequate to predict extremophilic behavior. Amino acid composition in thermophilic proteins have also been correlated to helix-coil transition parameters (Warren & Petsko 1995). This study suggested that the alpha helical frequencies of Val, Glu, His, Cys and Asp are decreased while those of Tyr, Gly and Gln are increased in thermophilic proteins. Haney et al. (1999) have compared sequences of 115 proteins from *Methanococcus jannaschii* (growth temperature 85°C) with known sequences of mesophilic

Methanococcus species. Their analysis indicated that composition of Ser, Asn, Gln, Thr, and Met are reduced in thermophiles while those of Ile, Arg, Glu, Lys, and Pro are increased. This analysis also suggested that the substitutions of uncharged polar residues, Ser, Thr, Asn, and Gln, by other residues are preferred. Thopmson and Eisenberg (1999) have suggested loop deletion as a contributing factor for thermal stability through their analysis of 20 different full genome sequences corresponding to organisms having different optimum growth temperatures.

Most of the understanding of structural factors contributing to the protein thermal stability has come from comparison of crystal structures of proteins isolated from mesophilic and thermophilic counterparts. Better hydrogen bonding networks, electrostatic interactions and internal packing have been suggested by various investigators as contributing factors to the stability of thermophilic proteins. These studies have been reviewed in Russell and Taylor 1995, Jaenicke and Bohm 1998, Ladenstein and Antranikian 1998. In the case of *Pyrococcus furiosis* rubredoxin (Day et al. 1992), *Thermotoga maritima* glyceraldehyde-3-phosphate dehydrogenase (Korndorfer et al. 1994) and *Pyrococcus furiosis* glutamate dehydrogenase (Yip et al. 1995), *Sulfolobus solgfataricus* indole-3-glycerolphosphate synthase (Hennig et al. 1995), increased number of ion-pairs and their arrangement into large ion-pair networks have been suggested to contribute to thermal stability. On the other hand in the case of *Thermus flavus* malate dehyrogenase (Kelly et al. 1993), *Thermoplasma acidophilum* citrate synthase (Russel et al. 1997) and *Bacillus stearothermophilus* triosephosphate isomerase (Delboni et al. 1995), increased compactness, shortening of loops, increased hydrophobicity and aromatic interactions, decreased flexibility of helical segments and subunit interfaces and stabilization of helical dipole have been suggested to contribute to thermal stability. Structure of Che Y protein from *Thermotoga maritama* showed no particular importance of any factor for its thermal stability (Usher et al. 1998). It was suggested that lowering of entropy of unfolding by the subtle rearrangement of several factors is responsible for thermal stability.

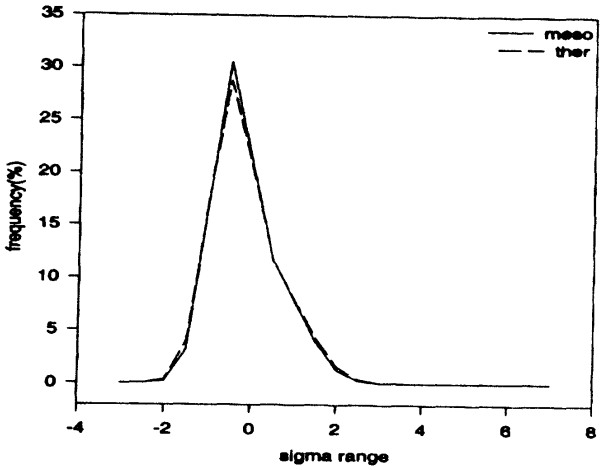

Figure 8 Frequency distribution of B'-factors for mesophilic and thermophilic protein structures

Table 2 *Values of parameters describing the summation of two Gaussian fit for thermophiles and mesophiles*

Parameter	K1	K2	B1	K3	K4	B2
Mesophiles	27.13	0.5914	5.942	9.48	0.2348	8.06
Thermophiles	25.38	0.5245	5.919	9.15	0.2133	8.14

Table 3 *p-values of amino acids for a bin size of 0.5*

Residue	Mesophiles	Thermophies
All	0.357	0.361
Ala	0.343	0.320
Cys	0.319	0.198
Asp	0.429	0.424
Glu	0.405	0.514
Phe	0.292	0.314
Gly	0.389	0.374
His	0.307	0299
Ile	0.296	0.306
Lys	0.429	0.446
Leu	0.287	0.340
Met	0.293	0.313
Asn	0.409	0.384
Pro	0.432	0.354
Gln	0.395	0.436
Arg	0.353	0.327
Ser	0.416	0.376
Thr	0.362	0.339
Val	0.307	0.294
Trp	0.268	0.291
Tyr	0.227	0.287

Facchiano et al. (1998) have analyzed helix stabilizing factors, like *N*-capping, charge-dipole interactions, hydrophobic interactions etc, and concluded that except for decrease in the number of β-branched residues other factors do not contribute significantly to stability of thermophilic proteins. Querol et al. (1996) have noted that better hydrogen bonding and better hydrophobic internal packing are among the most frequently quoted reason for thermal stability. Statistical analyses performed more recently addressing the relative importance of these two factors suggested that electrostatic interactions are more crucial than hydrophobic interactions (Vogt & Argos 1997, Vogt et al. 1997, Xiao & Honig 1999). Contradictory results were obtained when glyceraldehyde-3-phosphate dehydrogenase and glutamate dehydrogenase from mesophilic and thermophilic organisms were compared (Korndorfer et al. 1994 & Knapp et al. 1997). Karshikoff and Ladenstein (1998) concluded that packing density is not very different between mesophilic and thermophilic proteins.

These varied explanations prompted us to examine the possible correlation between the statistics of atomic displacement parameters and thermal stability. An overall increase in conformational rigidity is likely to be an additional factor contributing to the adaptation of proteins to extreme environments. Therefore, it is likely that the dynamical features of thermophilic and mesophilic proteins have significant and recognizable differences. 21 thermophilic protein structures determined at resolution better than 2.5 Å were chosen for this analysis (Parthasarathy & Murthy 2000a).

B′-factor Frequency Distribution

Figure 8 shows the frequency distribution of *B′*-factors for mesophilic and thermophilic proteins. The plots correspond to 30,960 amino acids for mesophiles and 10,469 amino acids for thermophiles. The *p*-values are 0.357 and 0.361 for mesophilic and thermophilic proteins, respectively. The six parameters characterizing the double Gaussian function are very similar for the two curves (table 2). Thus the mesophilic and thermophilic protein structures show similar *B′*-factor frequency

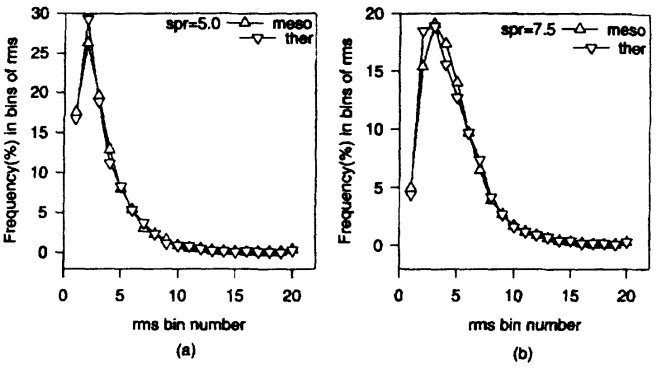

Figure 9 Frequency distribution of r.m.s. of *B′*-factors within spheres placed at C_α atomic positions for mesophilic and thermophilic proteins. for sphere radius of 5.0 Å (a), 7.5 Å (b) .

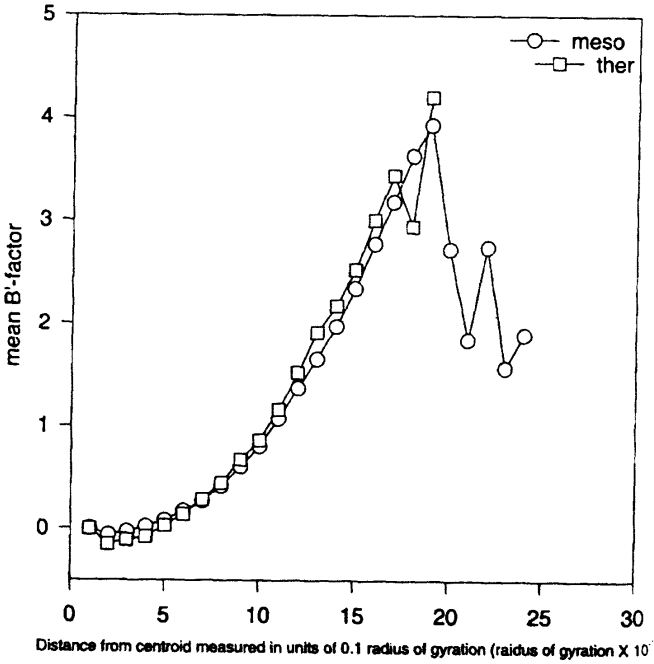

Figure 10 Plot of mean *B′*-factor in bins expressed as fractions of radius of gyration for mesophilic and thermophilic proteins.

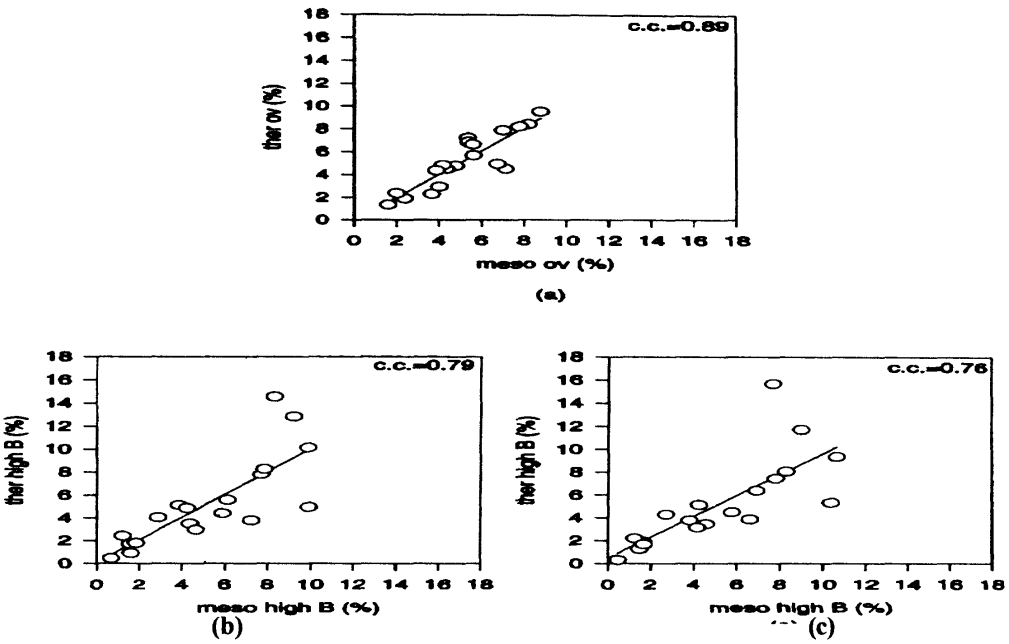

Figure 11 Scatter plots showing amino acid compositions in mesophiles and in thermophiles. (a) over all composition, (b) composition of residues with $B > \langle B \rangle + 0.50s(B)$, (c) composition of residues with $B > \langle B \rangle + 0.75s(B)$.

Table 4 *Amino acid composition of residues with high B-values. The first four columns give the number of residues and its composition (%) in mesophiles and thermophiles, respectively. Columns 5 and 6 represent the number and composition (%) of residues with* $B > \langle B \rangle + 0.5s(B)$ *in mesophiles. Column 9 and 10 represent these numbers for residues with* $B > \langle B \rangle + 0.75s(B)$. *Column 7,8,11 and 12 give the values corresponding to column 5,6,9 and 10 for thermophiles. Large differences between mesophiles and thermophiles are highlighted.*

	1	2	3	4	5	6	7	8	9	10	11	12
Ala	2723	8.80	997	9.52	272	7.69	87	7.81	227	7.80	71	7.43
Cys	439	1.42	67	0.64	33	0.93	0	--	27	0.93	0	---
Asp	1740	5.62	596	5.69	277	7.83	92	8.26	241	8.28	77	8.05
Glu	1659	5.36	763	7.23	294	**8.31**	162	**14.54**	224	**7.70**	150	**15.69**
Phe	1243	4.01	308	2.94	56	1.58	19	1.71	48	1.65	18	1.88
Gly	2552	8.24	881	8.42	350	9.89	113	10.14	310	10.65	95	9.34
His	732	2.36	199	1.90	57	1.61	10	0.90	43	1.48	12	1.26
Ile	1666	5.38	719	6.87	102	2.88	45	4.04	79	2.71	41	4.29
Lys	1728	5.58	695	6.64	326	**9.21**	143	**12.84**	262	**9.00**	112	**11.72**
Leu	2416	7.80	860	8.21	137	3.87	57	5.12	123	4.23	49	5.13
Met	617	1.99	248	2.37	43	1.22	27	2.42	36	1.24	21	2.20
Asn	1480	4.78	500	4.78	217	6.13	62	5.57	201	6.91	61	6.38
Pro	1373	4.34	471	4.50	208	5.88	49	4.40	168	5.77	43	4.50
Gln	1127	3.64	240	2.29	155	4.38	39	3.50	133	4.57	33	3.45
Arg	1293	4.18	507	4.84	150	4.24	54	4.85	111	3.81	36	3.77
Ser	2215	7.15	474	4.53	351	**9.92**	55	**4.94**	302	**10.38**	51	**5.33**
Thr	2091	6.75	519	4.96	256	**7.23**	42	**3.77**	192	**6.60**	37	**3.87**
Val	2176	7.03	826	7.89	165	4.66	33	2.96	121	4.16	30	3.14
Trp	491	1.59	141	1.35	24	0.68	5	0.45	13	0.45	3	0.31
Tyr	1199	3.87	458	4.37	66	1.86	20	1.80	49	1.68	16	1.67

distribution. Table 3 compares the *p*-values for individual amino acids between these two sets. It can be seen that Glu, Lue, Tyr and Gln have higher *p*-values in thermophiles when compared to mesophiles while Cys, Asn, Pro, Arg and Ser have lower *p*-values.

Comparison of Packing Densities in terms of B-Values

It is expected that in the well-packed interior of proteins, atoms will have strongly correlated displacements. It may be anticipated that the scatter (r.m.s value) of *B*-values in spheres around Cα will be lower in structures with better packed interiors as the correlation between atomic displacements of neighboring atoms in such structures will be stronger. Figure 9 shows the frequency distribution of r.m.s. in these spherical regions for mesophiles and thermphiles, for radii of 5.0 Å and 7.5 Å, respectively. There is no significant change in peak position or frequency in bins between mesophilic and thermophilic proteins, suggesting that the *B*-value distribution in spheres around Cα atoms does not reflect packing differences between mesophilic and thermophilic proteins.

Radial Increase of B′ from Centroid of Proteins

Comparison of the increment in *B*-values (Bhaskaran & Ponnuswamy 1988) between mesophiles and thermophiles from core of the protein to the exterior might also reflect differences in the packing of residues. The increment in *B′*-factors in spherical shells of radius corresponding to a specified fraction of the radius of gyration was examined. The slope of this incremental curve (figure 10) will be smaller in structures with better-packed interiors. It can be seen from figure 10 that there are no noticeable differences in the slopes. The fluctuation observed in mesophiles at outer edge may be due to loops extending out from the rest of the protein. These results agree with an earlier report that mesophilic and thermophilic proteins do not differ significantly with respect to packing interactions (Karshikoff & Ladenstein 1998). It is also possible that the accuracy of *B′*-factors is insufficient to reveal packing differences.

Frequency of high B-Value Stretches and Their Amino Acid Composition

The frequency of occurrence of consecutive residues with *B*-values greater than $\langle B \rangle + 0.5\sigma(B)$

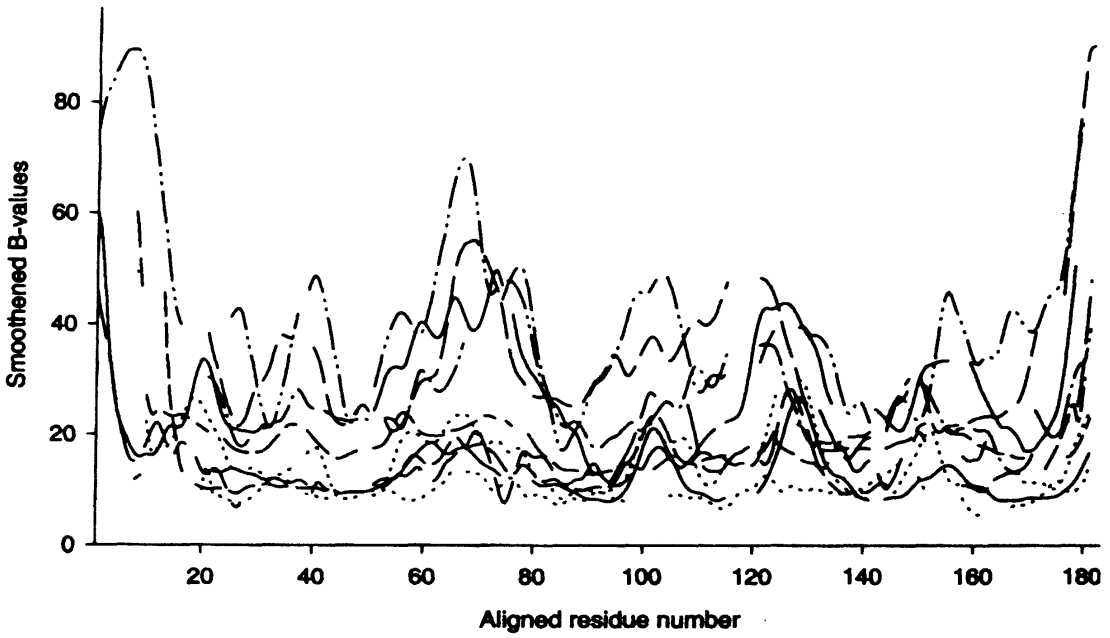

Figure 12 Plot of *B*-values of representative structures of hemoglobin smoothened over a window size of 5 against aligned residue number.

and of length 1 to 5, were counted for mesophilic and thermophilic proteins. No substantial difference in terms of the length distribution or frequency of amino acids is found in these stretches between thermophiles and mesophiles.

The scatter plot illustrating amino acid compositions of these segments is shown in figure 11. Table 4 lists the corresponding numbers. Figure 11a corresponds to over all amino acid composition and figure 11 b and c are plots corresponding to high *B*-value residues. As seen in Figure 11a and table 4, the overall amino acid composition is very similar in mesophiles and thermophiles (correlation co-efficient 0.89). In contrast, the correlation co-efficient between compositions of high B-value residues drops to 0.79 and 0.76 for cut off values of $<B>+0.5\sigma(B)$ and $<B>+ 0.75\sigma(B)$, respectively. The percentage composition of Glu residues in high *B*-value regions is nearly twice and that of Lys is nearly one-and-a-half times in thermophiles when compared to those of mesophiles. In contrast, the percentage compositions of Ser and Thr in high *B*-value regions of thermophiles are decreased by half (table 4). These are also reflected in larger *p*-values (table 3) for Glu, Lys and smaller p-values for Ser and Thr in thermophiles.

It is likely that replacement of Ser and Thr residues occurring in high *B*-value segments of mesophiles by other residues such as Ala leads to improvement of thermal stability. Mutations of Ser to Ala and Thr to Ala in mesophilic lactate dehydrogenase has been reported to enhance the stability of the enzyme by 20°C or more when compared to the wild type enzyme (Kotik & Zuber 1993). Similarly, mutations in streptococcal protein G β1 domain (Malakauskas & Mayo 1998) involving Ser and Thr residues have also led to considerable increases in thermal stability suggesting that the conclusions drawn by the present analysis are likely to be significant.

Conservation of Protein Dynamics During the Course of Evolution

Protein function depends on both its structure and dynamics. Hence it is likely that the evolutionary pressures on proteins operate towards the preservation of not only the structure but also the associated dynamics. In order to estimate the extent to which this preservation of dynamics is realized, ADPs of three families of proteins, haemoglobin, trypsin and triosephosphate isomerase were examined (Lenin et al. 2000).

Ten representative structures of hemoglobin (maximum sequence similarity 28%) and structures of all the native haemoglobin chains representing helical (α) class of proteins currently available in the PDB with resolution 2.0 Å or better were extracted from the Protein Data Bank (PDB).

Table 5 *Correlation coefficients between B-values of representative structures of hemoglobin*

	1ASH	1ECA	1FLP	1HLB	1ITH	1PBXA	1PBXB	2HBG	1LHB	3SDHA
1ASH	1.00									
1ECA	0.28	1.00								
1FLP	0.33	0.30	1.00							
1HLB	0.21	0.22	0.55	1.00						
1ITH	0.22	0.14	0.41	0.41	1.00					
1PBXA	0.53	0.28	0.42	0.29	0.10	1.00				
1PXAB	0.37	0.30	0.79	0.64	0.53	0.49	1.00			
2HBG	0.47	0.17	0.29	0.39	0.34	0.17	0.16	1.00		
2LHB	0.48	0.19	0.34	0.54	0.36	0.37	0.43	0.77	1.00	
3SDHA	0.40	0.37	0.30	0.50	0.28	0.35	0.23	0.62	0.62	1.00

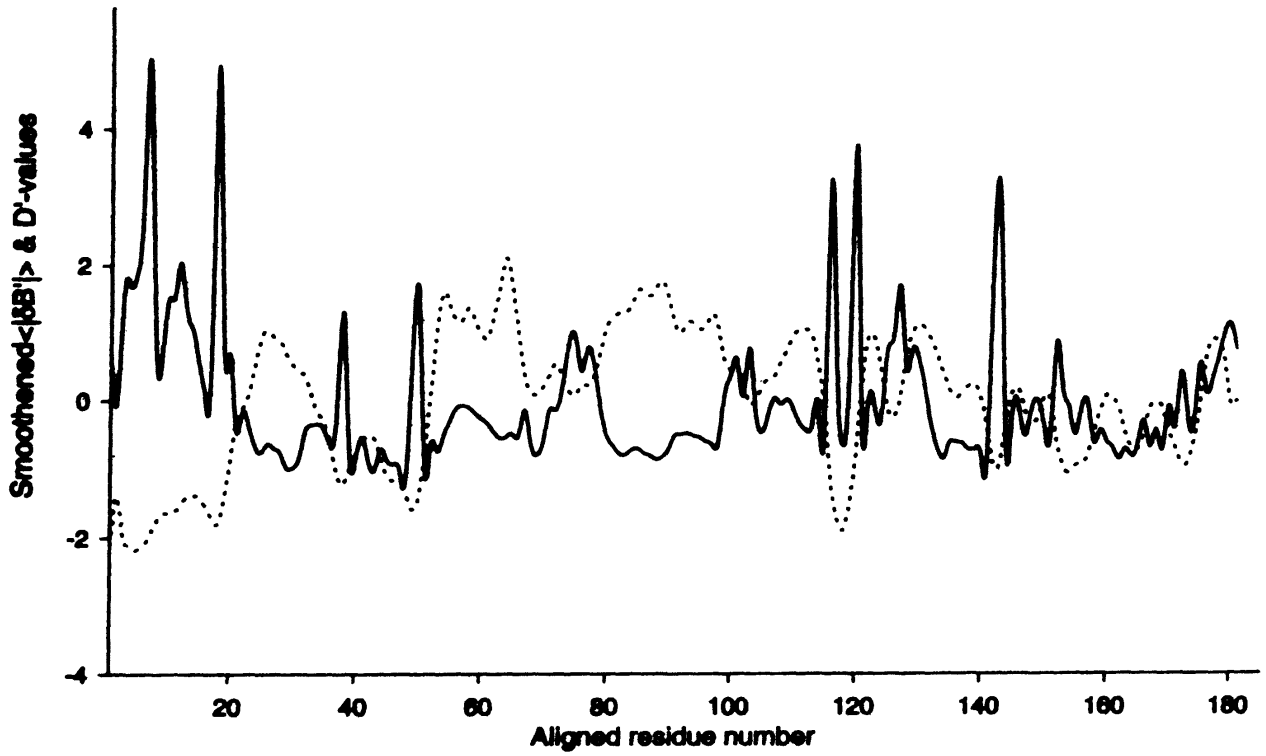

Figure 13 Profiles of smoothened $< |dB'| >$ values and smoothened D-values of representative hemoglobin structures

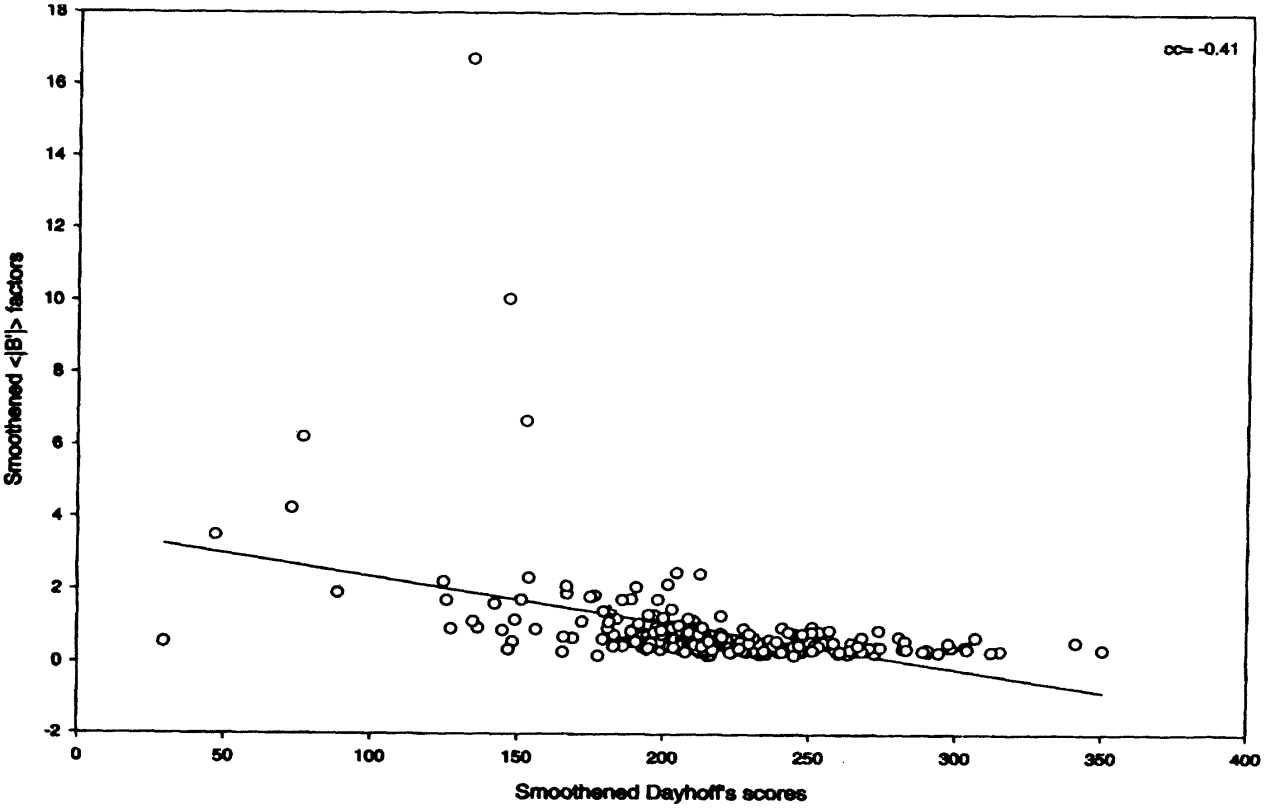

Figure 14 Scatter plot of pair-wise correlation coefficients of B-values and pair-wise sequence identity for trypsin structures

Similarly, structures of trypsin representing the β class and triose phosphate isomerase (TIM) representing the α/β class of proteins were selected. The resolution was better than or equal to 2.2 Å for trypsin structures and 2.8 Å for TIM structures. The maximum sequence identity between any two structures were 77% and 84% for trypsin and TIM, respectively.

Analysis of Representative Hemoglobin Structures

Multiple Alignment of Protein Sequences

Multiple sequence alignment was achieved using the PileUp program available in the Wisconsin GCG Package. For all sequence alignments, the standard scoring matrix supplied by the PileUp program was used. Sequence alignments of 10 representative hemoglobin structures were not meaningful due to their very low sequence similarity. To overcome this limitation, alignments were carried out after including a large number of haemoglobin sequences from other sources to establish a chain of relationships that might link these ten unrelated sequences. Multiple alignment was achieved with a gap creation penalty of 12, a gap extension penalty of 4 and an end-weight of 0.25. The correlation co-efficient of *B*-values at aligned positions between pairs of these structures were evaluated as,

$$\Sigma_i(B_{1i}-\langle B_1\rangle)\ (B_{2i}-\langle B_2\rangle) / \{\Sigma_i(B_{1i}-\langle B_1\rangle)^2\ \Sigma_i(B_{2i}-\langle B_2\rangle)^2\}^{1/2},$$

where B_{1i} is the *B*-value at position i in protein 1 and B_{2i} is the *B*-value at the equivalent position in protein 2. $\langle B_1\rangle$ and $\langle B_2\rangle$ are the averages of B-values in structures 1 and 2, respectively. The plot of smoothened B-values (window size 5) versus aligned sequence number is shown in figure 12. All the ten structures exhibit large humps approximately between aligned residue numbers 51 and 85 and between 111 and 135.

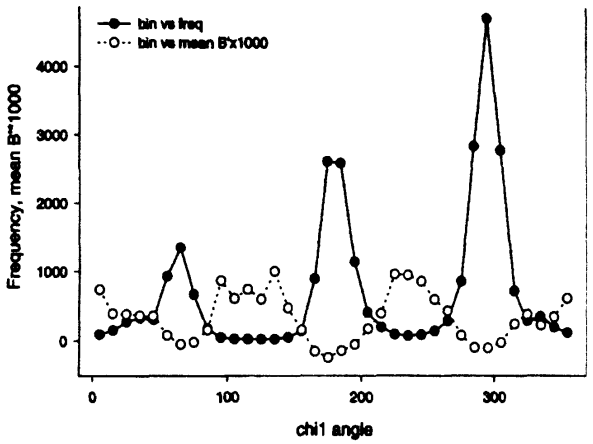

Figure 15 Frequency distribution of χ^1 and dependence of B-factor on χ^1

Omitting positions with gaps, the correlation coefficients between the B-values of the aligned residues were determined for each of the 45 unique pairs of structures. The correlation coefficients are shown in table 5. It can be seen that the correlation co-efficients between the B-values of most pairs are significant. This suggests that the dynamics of certain segments have been conserved even when the sequence similarity has almost disappeared.

Correlation Between B-values and Sequence Similarity

Dayhoff's score (Dayhoff et al. 1978) measures the total amino acid invariance at a particular residue. Therefore, it is likely to be correlated with a parameter that measures the total *B'*-variation at that particular residue. In order to measure the *B'*-variation at that position, a normalized parameter $\langle |\delta B'| \rangle$ was defined. $|\delta B'|$ refers to the absolute values of the differences in the *B'*-values between pairs of residues at a particular position and $\langle |\delta B'| \rangle$ is the average of such differences over all possible pairs. Figure 13

Table 6 *Frequency of residues in Ramachandran in bins and respective average B'-factors in brackets*

4378 (-0.19)	6220 (-0.06)	733 (0.23)	16 (0.85)	138 (0.25)	91 (0.25)
709 (0.04)	1652 (-0.06)	55 (0.94)	47 (0.79)	39 (0.46)	11 (1.87)
309 (0.20)	1253 (0.32)	16 (0.65)	453 (0.23)	722 (0.38)	13 (0.00)
167 (0.32)	9611 (-0.09)	3382 (0.12)	20 (0.33)	348 (0.56)	22 (0.26)
36 (0.35)	110 (0.42)	64 (0.83)	28 (0.64)	34 (1.65)	9 (0.8)
244 (0.00)	243 (0.21)	15 (0.59)	75 (0.17)	177 (0.17)	114 (0.02)

shows the profiles of smoothened D'-values and $<|\delta B'|>$ vs. the aligned residue number. The profiles of the two curves indicate that there is a degree of anti-correlation between the two parameters.

Analysis of All Native Structures of Haemoglobin

The plot of smoothened *B'*-values of the structures against the aligned sequence number retained the general features of corresponding plot of the representative structures, but that they were shifted to the left by a few residues. This shift is probably due to the reduction in the total alignment length resulting from the greater similarity between the structures and fewer gaps needed for the alignment. The minor humps are amplified in this case.

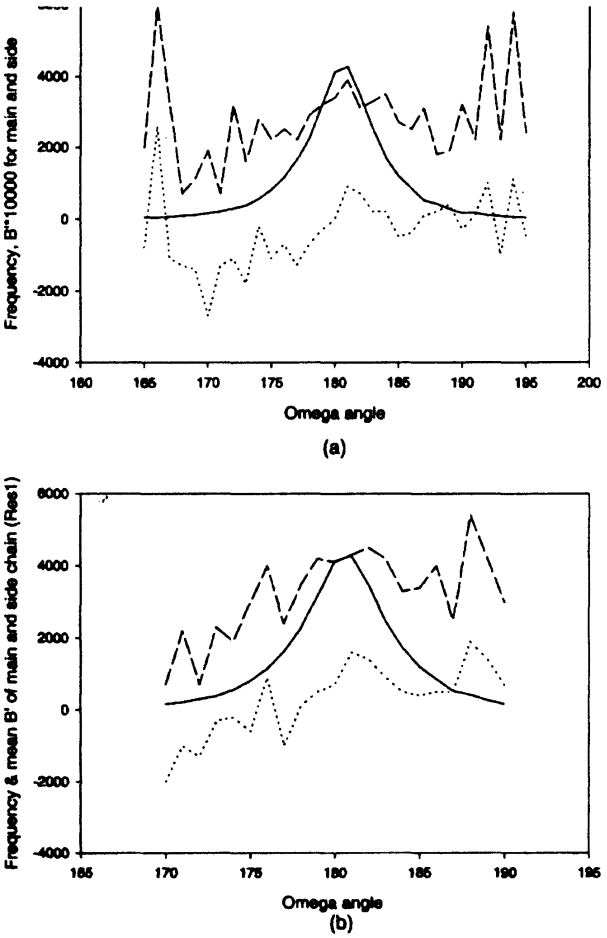

Figure 16 Variation of mean *B'*-factor as a function of peptide plane distortion ω. (a) main (.....) and side (_ _ _) chain atoms of the residue preceding and (b) main (.....), and side (_ _ _) chain atoms of the residue following the peptide bond

Analysis of Structures of Trypsin and Triosephosphate Isomerase

Globins are α-helical proteins. In order to examine the validity of these results in other classes, a set of structures representing β-sheet proteins (trypsins) and a representative set for α/β proteins (TIMs) were selected and the analyses were repeated.

The sequences of trypsins and TIMs were aligned using the PileUp program with gap creation penalty of 12 and gap extension penalty of 4. In both cases, the smoothened *B'*-value vs. aligned residue number plots for all the structures were almost identical. The correlation coefficients between the *B*-values were high for all the pairs. There was a small negative correlation between $<|\delta B'|>$ and Dayhoff's scores suggesting only a weak link between sequence changes and the corresponding changes in flexibility. Figure 14 is such a plot for trypsin structures. Interestingly, the sequences of trypsin chosen clustered into two distinct classes with low Dayhoff's score between members of one class and those of the other. However, for sequence pairs of high and low similarities, the correlation coefficients were similarly scattered suggesting that the changes in sequence do not result in substantial changes in flexibility. Similar results were obtained for TIM structures.

The closely similar results in the three distinct classes of proteins examined suggest that polypeptide flexibility is retained to a large extent in the course evolution. Thus the relationship between sequence and flexibility has a similar degeneracy as the relationship between sequence and structure. However, it is worth noting that the observed correlation might not be a direct consequence of selection. It is possible that the evolutionary pressure conserves the three-dimensional structure, which in turn, dictates the flexibility. This conservation of protein flexibility during the course of evolution is, however, consistent with the generally accepted importance of flexibility to protein function.

Relation between *B*-Values and Protein Conformational Properties

The native structure of a protein is likely to correspond to the global minimum of conformational

energy. In the native structure, there will be no net force on the atoms at their equibrilium position. The extent of atomic displacement from their position depends on the energy gradient, which is a function of the local environment. It is therefore possible that the atomic displacement parameter (*B*-values) might be influenced by local conformation.

The following angular parameters define protein conformation. χ^1 defining protein side chain conformation is the angle between the planes containing N - $C\alpha$ – $C\beta$ and $C\alpha$ – $C\beta$ - $C\gamma$ ($O\gamma$ for Ser and $C\gamma 1$ for Ile) atoms. ω angle representing the deviation of the peptide geometry from a planar configuration is the torsional angle $C\alpha_i$ – C'_i –N_{i+1} – $C\alpha_{i+1}$ (where i and i+1 represent consecutive residues). 0° and 180° in ω correspond to ideal cis and trans peptide geometries, respectively. Ramchandran angles ϕ and φ representing the main chain conformation at each $C\alpha$ atom are the angles C'_{i-1} - N_i - $C\alpha_i$ - C'_i and N_i - $C\alpha_i$ - C'_i – N_{i+1}, respectively. Variation of mean *B*'-factor was monitored as a function of χ^1 angle by grouping side chains into bins of 10°. Sampling

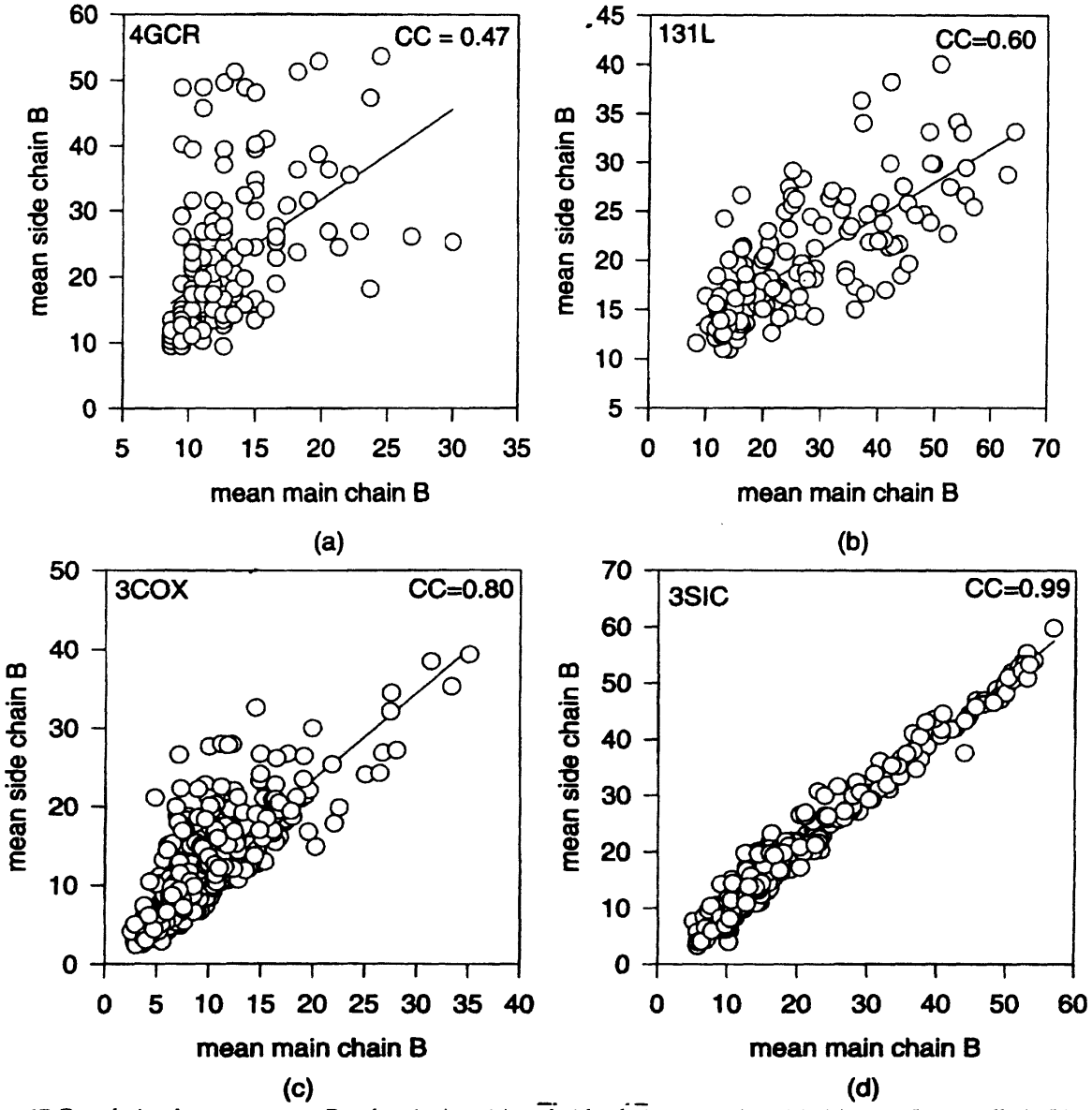

Figure 17 Correlation between mean Bs of main (x-axis) and side chain atoms (y-axis), (a) 4gcr. (b crystallin), (b) 131L (T4 lysozyme mutant), (c) 3cox (cholesterol oxidase), (d) 3sic (substilisin inhibitor complex)

was at 1° intervals in ω. Frequencies were counted in 60° bins of ϕ and φ.

Degrees of Flexibility in Main and Side Chain Conformations

Figure 15 shows the frequency distribution of side chain rotomers and their mean *B*′-factors. The χ^1 distribution follows features that have been reported earlier. The peaks corresponding to the preferred rotamers have widths at half maximum of ~25°. The frequencies for intervals close to the preferred torsion angles are large and hence the estimated *B*′-factors are reliable. The *B*′-factor has clear minima at the preferred conformations. It might also be noted that the most favored torsion angle is less than 300°. The minimum in mean *B*′-factor also occurs at this value. Similar observations have been made earlier (Dunbrack & Karplus 1993).

Table 6 lists the frequencies of residues occurring in different bins of Ramchandran plot (Ramakrishnan & Ramachandran 1965). The largest occupancy is found for the helical regions in the third quadrant (negative φ, negative ϕ) and for strand regions in the second quadrant (positive φ, negative ϕ) of the map. The *B*′-factors associated with the atoms N_i, $C_i\alpha$, C_i',N_{i+1} of all residues falling in the bins were averaged. The mean *B*′ factor of all Cα atoms was subtracted from this average. The values shown in table 6 correspond to these values. It is evident that the mean *B*′-factor is low for all cells with high occupancy. Since the frequency of occurrence depends on the energy corresponding to the particular cell, the flexibility is linked to the energy as observed for χ^1. Similar observations were made for the *B*′-factors associated with the Cα atom alone. Thus the occurrence of a particular pair of Ramachandran angle has a small but perceptible effect on the flexibility of the local segment of the polypeptide.

Changes in Flexibility Induced by Rotation of the Peptide Plane

Figures 16a and 16b show the frequency distribution of ω angle, representing out of plane deviation of the trans peptide unit about the C′-N bond along with the variation of the mean *B*′-factor as a function of ω for the preceding and

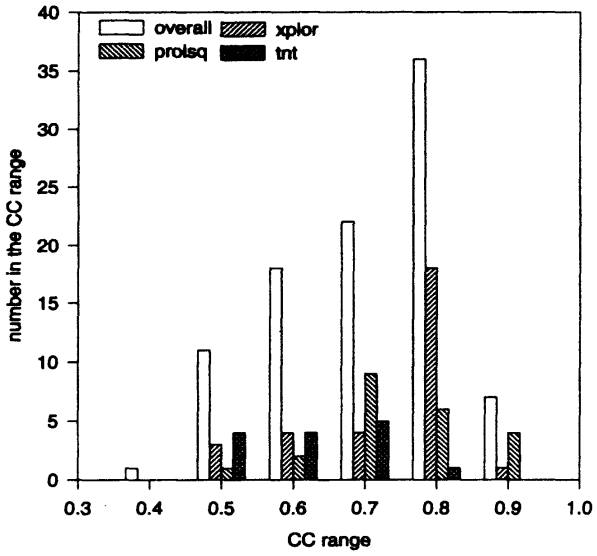

Figure 18 Histogram illustrating frequency distribution of CC

succeeding residues, respectively. The highest frequency does not correspond to the ideal trans geometry, but for ω= 179° (MacArthur & Thornton 1996). The distribution is nearly symmetrical about the preferred angle. The mean *B*′-factor appears to monotonically increase with increasing ω in the range 170° – 190° (figure 16). It implies that the peptide group becomes more flexible when ω increases from the value corresponding to the ideal trans geometry. Ramakrishnan and Balasubramanian (1972) have shown that the area allowed in the Ramachandran diagram for non-glycyl residues reduces with increasing ω, in the range 170° – 190°. It is therefore plausible that the reduction of *B*′-factor for non-planar distortions with ω < 180° is due to the larger allowed area (lack of close contacts) of the Ramachandran diagram for the corresponding ω.

Correlation Between B-values of Main Chain and Side Chain Atoms

The reported *B*-values of protein structures are influenced by the constraints and restraints employed for refinement. It is possible to study the effects of restraints on reported *B*-values by analysing the correlation coefficients (CCs) between the mean *B*-values of main chain and side chain atoms (Parthasarathy & Murthy 1999). We grouped the 95 structures on the basis of the package used for their refinement. Thirty of these structures

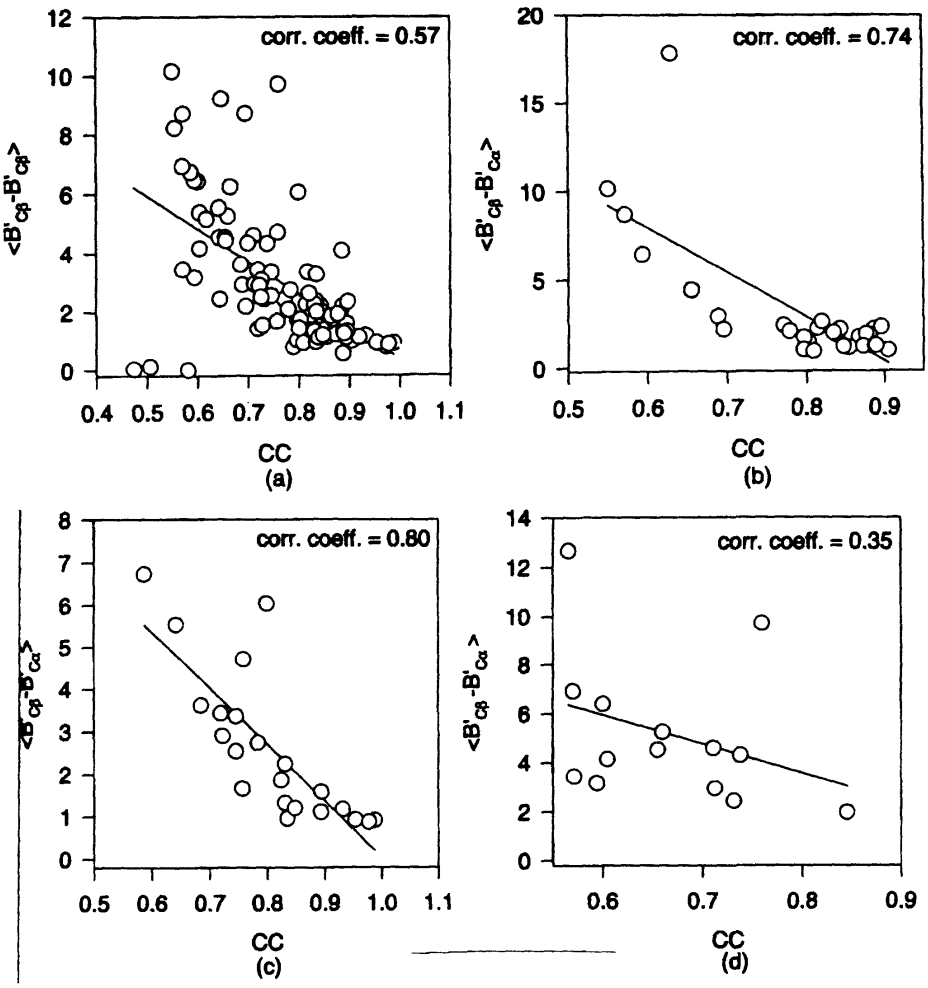

Figure 19 Dependence of the CC on mean difference in the B'-factors of Cα and Cβ atoms (a) all structures, (b) structures refined using X-PLOR, (c) structures refined using PROLSQ, (d) structures refined using TNT

Table 7 Length distribution of stretches with high B-values at Cα atoms in structures with low CC (0.5 - 0.6) and high CC (0.85 -0.95) values. The number in brackets are the percentage of total residues that belong to these segments.

Length	Low CC	High CC
1	213 (6.4)	37 (1.4)
2	78 (4.7)	16 (1.2)
3	26 (2.4)	12 (1.3)
4	18 (2.2)	15 (2.2)
5	7 (1.1)	6 (1.1)
6	7 (1.3)	9 (2.0)
7	11 (2.3)	3 (0.8)
8	2 (0.5)	10 (3.0)
9	2 (0.5)	3 (1.0)
10	2 (0.6)	3 (1.1)
11	0 (0.0)	4 (1.6)
12	1 (0.4)	3 (1.3)
14	0 (0.0)	1 (0.5)
15	0 (0.0)	1 (0.6)

were refined using X-PLOR (Brunger et al. 1987), 22 by PROLSQ (Hendrickson & Konnert 1980) and 14 by TNT (Tronrud et al. 1987). The rest of the 21 structures were refined using either a combination of these packages or by using some other protocol.

In order to understand the difference in CCs observed for structures refined using the same package, dependence of CCs on mean difference in the *B'*-factors of Cα and Cβ atoms was examined.

For each protein, the correlation coefficient between the main chain and the side chain thermal parameters was computed as,

$$CC = \frac{\Sigma (B_{main} - \langle B_{main} \rangle)\,(B_{side} - \langle B_{side} \rangle)}{[\Sigma (B_{main} - \langle B_{main} \rangle)^2\ \Sigma (B_{side} - \langle B_{side} \rangle)^2]^{1/2}}$$

The summation is over all the residues except glycine.

For 61% of the structures studied, the CC values fall between 0.7 and 0.9. The mean value of CC is 0.76. Figure 17(a-d) shows representative examples of scatter plots illustrating the correlation between the mean *B*-values of the main chain and the side chain atoms for individual proteins. The lowest value of CC, 0.47, is exhibited by γ-crystallin (4GCR) and the highest value, 0.99, is observed for substilisin-inhibitor complex (3SIC). The contrast between the plots for these proteins (figure 17a &d) is striking.

Effects of Experimental Temperature, Protein Size and Crystal Packing on CC Values

It is possible that the observed *CC* values are influenced by the temperature during the data collection, molecular weight of the protein and crystal packing. These aspects were examined by analysing the CC values of lysozyme whose structure has been determined at different temperatures (Kurinov & Harrison 1995). The *CCs* for all these structures are around 0.57 and vary between 0.53 and 0.63. There is no significant influence of experimental temperature on CC values.

For globular proteins, the surface to volume ratio reduces with size. Therefore the percent of residues buried inside a folded protein is likely to increase with the protein size. The degree to which the thermal parameters of the side chain atoms could be different from those of the main chain atoms might depend on their exposure to solvent. Therefore, there might exist a correlation between the CCs and the molecular weight. Another parameter to which CCs are likely to be related is the Matthews coefficient (Matthews 1968). Low Matthews coefficient represents closely packed structures where the thermal parameters are less likely to be independent. No significant dependence of CC on molecular weight or Matthews coefficient was observed.

Effects of Program Used for Refinement

Structures were segregated into four groups based on the program used for refinement; X-PLOR, PROLSQ and TNT. Frequencies in CC ranges for each of the groups were counted. The results are shown as histograms in figure 18. The most frequent CC range is 0.8-0.9 (36 out of 95). The distribution in different CC ranges for proteins refined using X-PLOR resembles the distribution for all the structures. Structures refined with PROLSQ have a similar number falling in the CC ranges 0.7-0.8 and 0.8-0.9. On the contrary, the structures refined using TNT show uniform frequency over the CC range 0.5-0.8. Thus, the program used for refinement influences the spread of CC values. The very high CC value observed in a few cases (3SIC, substilisin-inhibitor complex, figure 17d), could reflect incomplete refinement.

Dependence of CC on the Restraints Used for B-Value Refinement

The mean size of the differences between the *B*-values of *Cα* and *Cβ* atoms is likely to reflect the differences in refinement strategies and packages. If this value is constrained to small values in the refinement protocol, it will lead to large CC as well as to long stretches of consecutive C*α* atoms with comparable *B*-values. Figure 19(a-d) illustrates the relationship between CC values and the mean difference between *B'*-factors of Cα and Cβ atoms for all the structures (figure 19a) and for structures refined using X-PLOR (figure 19b), PROLSQ (figure 19c) and TNT (figure 19d). It is obvious that CCs are correlated to the size of the differences. $\langle B'_{C\beta}-B'_{C\alpha}\rangle$ appears to be less tightly related to CC in structures refined with TNT when compared to those refined with X-PLOR or PROLSQ.

Frequency Distribution of High B-values Stretches

The occurrence of stretches of high *B*-values, $B(C\alpha) > \langle B\rangle + 0.5\ \sigma(B)$ were counted. The frequency distributions of these high *B*-value stretches were calculated for two sets of proteins: with low CC (0.5-0.6) and with high CC (0.85-0.95). Table 7 lists the distribution. The fraction of residues that occur with high B values is about the same in proteins with low and high correlation coefficients (739 out of 3303 in the low CC proteins and 612 out of 2691 in high CC proteins occur

with high B-value, amounting to 22.4% and 22.7%, respectively). In spite of this comparable fraction of high B residues, very long stretches of high B-values are not observed in proteins with low CC when compared to proteins with high CC. In contrast, short stretches of high Bs occur more frequently in proteins with low CC.

More frequent observation of longer stretches of consecutive high B-values (table 8), smaller difference between the B'-factors of Ca and Cb atoms and high CCs are generally found simultaneously. It is likely that these structures have been refined with smaller weights to the X-ray term relative to weights for stereochemical constraints or B-values when compared to structures with lower CCs.

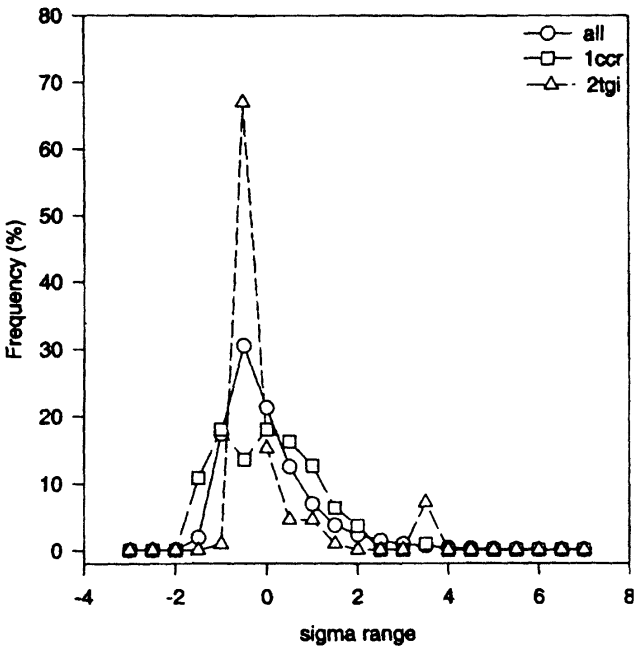

Figure 20. Observed B'-factor frequency distribution in bins of size 0.5. over all distribution (o); distribution for rice ferricytochrome *c* (1CCR, ◊); distribution for transforming growth factor β2 (2TGI, Δ).

Possible Validation Tool for the Reported B-values

Overview of Currently Available Structure Validation Methods

Crystallographic residual, the R-factor ($R = \Sigma|F_o - F_c| / \Sigma F_o$, where F_o is the observed amplitude and F_c is the amplitude calculated from the model), has conventionally been used as an index of the agreement of the derived protein model with the X-ray diffraction data. Recently, introduction of the free R-factor has improved the reliability of structural refinements (Brunger 1992, Kleywegt & Brunger 1996). The free R-factor is the crystallographic residual (R) for reflections that are not included in the refinement. These methods, however, are only global indicators of correctness. With advances in methods of crystallization, data collection and refinement techniques, there is a rapid increase in the rate of structure determination. Validation of structural parameters and recognition of errors in these protein models is a crucial task (Branden & Jones 1990, Janin 1990). During protein refinement, both geometrical (bond lengths, bond angles, planarity of aromatic, peptide and side chain guanidino, amido and carboxylate groups and chirality of tetrahedral carbon atoms) and conformational (backbone and side chain torsion angles, ring puckerings and van der Walls interactions) restraints are used. The geometrical parameters are usually scattered around a preferred value (unimodal) while conformational properties are scattered around several distinct values (multimodal). While the geometrical properties are restrained greatly during the refinement, the conformational properties are, generally, not restrained (except in torsion angle refinement option in XPLOR) to the same extent. The errors in geometrical properties might be estimated on the basis of the knowledge derived from studies on small molecules. The expected distributions of conformational properties have been derived directly from protein structures. Thus, the deviations from the expected distribution of conformational properties are sensitive indicators of errors in the reported models.

A variety of other methods to recognize errors in protein structures have also been suggested. They include methods independent of X-ray data, such as three-dimensional profiles (Luthy et al. 1992), patterns of non-bonded interactions (Colovos & Yeates 1993), methods based on knowledge of mean force fields (Sippl 1993), protein solvation properties (Holm & Sander 1992)

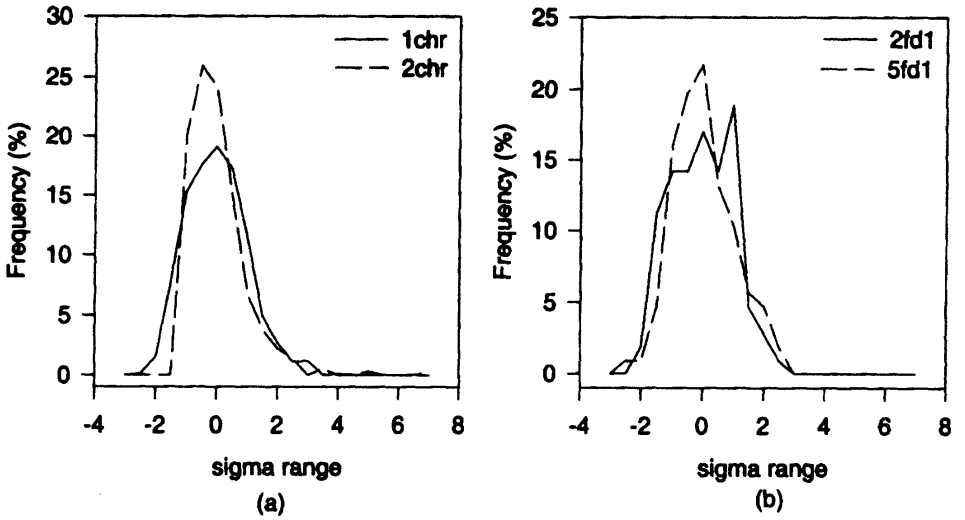

Figure 21. Observed B'-factor frequency distribution for initial erroneous structures (solid lines) and subsequently corrected structures (broken lines). (a) Chloromuconate cycloisomerase (1CHR & 2CHR), (b) Ferridoxin (2FD1 & 5FD1).

and directional properties of atomic contacts (Vriend and Sander 1993). Hooft et al. (1996) have compared the planarity of side chain groups in protein and small molecule structures. Validation packages carrying out a variety of checks on the co-ordinates, report abnormalities and suggest outliers are available. The popular ones are PROCHECK (Laskowski 1993), WHAT IF (Vriend 1990), SQUID (Oldfield 1992) and PROVE (Pontius et al. 1996). The first three programs carry out various conformational checks by comparing the model and a library derived from a database of selected structures. They detect wrong tracing of structures and abnormal torsion angles. PROCHECK database has been derived from 119 non-homologous high-resolution protein structures. The database used in WHAT IF is updated periodically. SQUID database can be generated using any set of high-resolution structures using PDBSEL program of CCP4 package. SQUID and WHAT IF also inspect hydrogen bonding pattern of His, Asn and Gln residues. PROVE stands for, PROtein Volume Evaluation and calculates atomic and residue volumes and makes comparison with a dictionary of standard values derived from 64 high-resolution structures using Voronoi method. In all these methods Ramachadran Map (Ramakrishnan & Ramachandran 1965, Kleywegt & Jones 1996) provides a simple and sensitive method to detect errors in back bone torsion angles and find special structural features which may be important for maintaining the local structure and function. Reliability of these validation methods have also been analysed using protein structures determined at resolution 1.2 Å or better (EU 3-D Validation Network, 1998). While several options are available in these validation packages for geometrical and conformational properties of protein models, no convenient option exits for checking the reported B-values. WHAT IF analyses the differences in the B-values of bonded atoms and gives warning if they are unusual. SQUID performs analysis of anisotropic B-values.

In this context, we have proposed the frequency distribution of B-values as a validation tool for the reported B-values (Parthasarathy & Murthy 1999). An estimate of the deviation of observed B'-factor frequency distribution of protein structure from the over all observed (or fitted) frequency distribution might be an index useful for validating the B-values. We have also compared B'-factor frequency distribution of structures that were reported wrongly earlier and have subsequently been corrected.

Frequency Distribution of B'-factor as a Possible Validation Tool

The B'-factor frequency distribution is a general characteristic of protein structures irrespective of their size, oligomerization, crystal packing interactions and experimental temperature. The characteristic double Gaussian fit was found to be good for most of the 95 structures examined although the refinement protocols followed for these structures were different. The B'-factor frequency distributions of individual proteins were compared to the overall fitted distribution. The average correlation coefficient between distribution for individual proteins and that of the overall distribution is 0.97 with a r.m.s. value of 0.03. Among the 95 structures analysed, the deduced parameters describing the double Gaussian fit were very different for 1CCR (ferricytochrome c; Ochi 1983) and 2TGI (transforming growth factor; Daopin et al. 1992). Their distributions also had low agreement with the overall distribution. Figure 20 shows the superposition of frequency distribution of these two structures with the over all fitted distribution. The agreement is poor. The correlation coefficients are 0.84 and 0.80, respectively, for 1CCR and 2TGI, compared to the average value of 0.97. These proteins lack the hydrophobic core usually found in the structures of other proteins. Visual inspection of these structures showed that 1CCR is loosely wound around the heam group while 2TGI has most of its side chain groups accessible to solvent. The original investigators also have noticed the special dynamic feature of the 2TGI structure. These observations suggest that the B'-factor frequency distribution could highlight protein structures with unusual dynamics.

Models of a few proteins have been wrongly built and reported and have subsequently been corrected. Choloromuconate cycloisomerase and Ferridoxin from *A.vinelandii* are two examples of such cases. The B'-factor frequency distribution for the initially reported structures (1CHR, Hoier et al. 1994, 2FD1, Ghosh et al. 1982) and later corrected versions (2CHR, Kleygwegt et al. 1996, 5FD1, Stout 1993) are shown in figure 21. The correlation between overall distribution and 2FD1 & 5FD1 are, 0.79 and 0.95, respectively. Values for 1CHR and 2CHR chloromuconate cycloisomerase are, 0.91 and 0.99, respectively. The distributions for only the corrected versions resemble the anticipated distribution. Thus it appears that the B'-factor frequency distribution could be used to qualitatively validate the B-values.

Conclusions

Comparative analysis of the B-values obtained by X-ray diffraction studies on different protein crystals is difficult due to the lack of adequate models for static and dynamic disorder, differences in the strengths of restraints employed and human subjectivity in the selection of refinement protocols. Inspite of these difficulties, protein flexibility appears to follow certain unique patterns. We have demonstrated possible applications of B-values to understand and interpret protein structure, evolution and dynamics. With the availability of atomic resolution structures and possibility of refining anisotropic displacement parameters, our understanding of protein dynamics and its relation to protein function will further be greatly enhanced.

Acknowledgements

This work was supported by the Department of Science and Technology and Council of Scientific and Industrial Research, India. SP thanks CSIR for financial assistance. Thanks are also due to Mr. V.M.S.. Lenin for help in some calculations related to dynamics of homologous proteins. We are grateful to Springer-Verlag, NewYork, Inc. for permission to reproduce figures 1 and 2, Cambridge University press, NewYork, for permission to reproduce figures 3-7, Oxford University press, Oxford, UK, for permission to reproduce figures 8-11, International Union of Crystallography for permission to reproduce figures 17-21, Current Science Association, Bangalore, India, for permission to reproduce figures 12-16.

References

Arias L M and Argos P 1989 Engineering protein thermal stability: Sequence statistics point to residue substitutions in (a-helices); *J. Mol. Biol.* **206** 397-406

Berstein F C, Koettzde T F, Williams G J B, Meyer E F Jr, Brice M D, Rodgers J R, Kennard O, Shimanouchi T, Tasumi M 1977 The protein Data Bank: A computer-based archival file for macromolecular structures; *J Mol Biol* **112** 535-542

Bhaskaran R and Ponnuswamy P K 1988 Positional flexibilities of amino acid residues in globular proteins; *Int. J. Peptide Protein Res.* **32** 241-255

Bohm G and Jaenicke R 1994 Relevance of sequence statistics for the properties of extremophilic proteins; *Int. J. Peptide Protein Res.* **43** 97-106

Branden C and Jones T A 1990 Between objectivity and subjectivity; *Nature* **243** 687-689

Brunger A T, Kuriyan J and Karplus M 1987 Crystallographic R-factor refinement by molecular dynamics; *Science* **235** 458-460

_____ 1992 Free R-value: a novel statistical quantity for assessing the accuracy of crystal structures; *Nature* **355** 472-475

Colovos C and Yeates T O 1993 Verification of protein structures: Patterns of nonbonded atomic interactions; *Protein Sci.* **2** 1511-1519

Daopin S, Piez K A, Ogawa Y and Davies D R 1992 Crystal structure of transforming growth factor-(2: An unusual fold for the superfamily; *Science* **257** 369-373

Dauter Z, Lamzin V S and Wilson K S 1997 The benefits of atomic resolution; *Curr. opnin. Stru. Biol.* **7** 681-688

Day M W, Hsu B T, Joshua-tor L, Park J, Zhou Z H, Adams W W and Rees D C 1992 X-ray crystal structures of the oxidised and reduced forms of hyperthermophilic archaebacterium *Pyrococcus furiosus; Protein Science* **1** 1494-1507

Dayhoff M O, Schwartz R .M and Orcutt B C 1978 In Atlas of protein sequence and structure National Biomedical Research Foundation, Washington DC 5 353-362

Delboni L F, Mande S C, Rentier-Delrue F, Mainfroid V, Turley S, Vellieux F M D, Martial J A and Hol G J 1995 *Protein Science* **4** 2594-2604

Dodson E J, Kleywegt G J and Wilson K 1996 Report of a workshop on the use of statiscal validators in protein X-ray crystallography; *Acta Cryst.* **D52** 228-234

Dunbrack R L Jr and Karplus M 1993 Backbone-dependent rotamer library for Proteins; *J. Mo.l Bio.l* **230** 543-574

EU 3-D Validation network 1998 who checks the checkers? Four validation toos applied to eight atomic resolution structures; *J. Mo. Biol.* **276** 417-436

Facchiano A M, Colonna G and Ragone R 1998 Helix stabilizing factors and stabilization of thermophilic proteins: an X-ray based study; *Protein Enging.* **9** 753-760

Ghosh D, Donnell S O, Fuery W, Robbins A H Jr and Stout C D 1982 Iron-sulfur clusters and protein structure of *Azotobacter ferridoxin* at 2.0Å resolution; *J. M.ol. Biol.* **158** 73-109

Haney P J, Jonathan H B, Buldak G L, Reich C I, Woese C R and Olsen G J 1999 Thermal adaptaion analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species; *Proc. Natl. Acad. Sci. USA* **96** 3578-3583

Hendrickson W A and Konnert, J.H. 1980 Incorporation of streochemical information into crystallographic refinement, in *Computing in Crystallography*, eds R Diamond and K Venkatesan, ch.13, pp. 13.01-13.26. (Bangalore: Indian Academy of Sciences)

Hennig M, Darimont B, Sterner, R., Kirschner, K. and Jansonius J N 1995 2.0Å structure of indole-3-glycerophosphate synthase from the hyperthermophile Sulfolobus solfataricus: possible determinants of protein stability; *Structure* **3** 1295-1306

Hobohm U and Sander C 1994 Enlarged representative set of protein structures; *Protein Sci.* **3** 522-524

Hoier H, Schlomann, M, Hammer, A, Glusker, J P, Carrel H L, Goldman A, Stezowski J J and Heinemann U 1994 Crystal structure of Chloromuconate cyclo isomerase from Alcaligenes eutrophus JMP134 (pJP4) at 3 Å resolution; *Acta Cryst.* **D50** 75-84

Holm L and Sander C 1992 Evaluation of protein models by atomic solvation preference; *J. Mol. Biol.* **225** 93-105

Hooft W W, Sander C and Vriend G 1996 Verification of protein structures: Side-chain planarity; *J. Appl. Cryst.* **29** 714-716

Jaenicke R and Bohm G 1998 The stability of proteins in extreme environments; *Curr. Opin. Strul. Biol.* **8** 738-748

Janin J 1979 surface and inside volumes in globular proteins; *Nature* **277** 491-492

_____1990 Errors in three dimensions; *Biochemie* **72** 705-709

Karplus P A, Schulz G E 1985 Prediction of chain flexibility in Protein; *Naturwissenschaften* **72** 212-213

Karshikoff A and Ladenstein R 1998 Proteins from thermophilic and mesophilic organisms essentially do not differ in packing; *Protein Engng.* **11** 867-872

Kelly C A, Nishiyama M, Ohnishi Y, Beppu T and Birkoft J J 1993 Determinants of protein thermal stability observed in the 1.9Å crystal structure of malate dehydrogenase from the thermophilic bacterium *Thermus flavus; Biochemistry* **32** 3913-3922

Kleywegt G J and Brunger A T 1996 Checking your imagination: applications of the free R value; *Structure* **4** 897-904

_____ and Jones A T 1996 Phi/Psi-chology: Ramachandran Revisited; *Structure* **4** 1395-1400

_____, Hoier H and Jones A T 1996 A re-evaluation of the crystal structure of chloromuconate cycloisomerase; *Acta Cryst.* **D52** 858-863

Knapp S, de Vos, W M, Rice D and Ladenstein R 1997 Crystal structure of glutamate dehydrogenase from the hyperthermophilic eubacterium Thermotoga maritima at 3.0 A resolution; *J. Mol. Biol.* **267** 916-932

Korndorfer I, Steipe B, Huber R, Tomschy A and Jaenicke R 1995 The crystal structure of holo-glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima* at 2.5Å resolution; *J. Mol. Biol.* **246** 511-521

Kotik M and Zuber H 1993 Mutations that significantly change the stability, flexibility and quaternary structure of the l-lactate dehydrogenase from *Bacillus megaterium; Eur. J. Biochem.* **211** 267-280

Kuhn P, Knapp M, Soltis S M, Ganshaw G and Bott R 1998 The 0.78 Å structure of a serine protease: *Bacillus lentus* subtilisin; *Biochemistry* **37** 13446-13452

Kurinov I V and Harrison R W 1999 The influence of temperature on Lysozyme crystals. Structure and dynamics of protein and water; *Acta Cryst.* **D51** 98-109

Ladenstein R and Antranikian G 1998 Proteins from hyperthermophiles: Stability and enzymatic catalysis close to the boiling point of water; *Adv. Biochem. Engng. Biotech.* **61** 37-85

Laskowski A 1993 PROCHECK: a program to check the stereochemical quality of protein structures; *J. Appl. Cryst.* **26** 283-291

Lenin V M S, Parthasarathy S and Murthy M R N 2000 Atomic Displacement Parameters (B-values) in homologous proteins: Conservation of dynamics structures; *Curr. Sci.* **78** 1123-1126

Libeu C, Kukimoto M, Nishiyama, M, Horinouchi, S and Adman E T 1997a Site-directed mutants of Pseudoazurin: Explanation of increased redox potentials from X-ray structures and from calculation of redox potential differences; *Biochemistry* **36** 13160-13179

_____ and Adman E T 1997b Displacement-parameter weighted co-ordinate comparison: I. Detection of significant structural differences between oxidation states; *Acta Cryst.* **D53** 56-77

Longhi S, Czjzek M and Cambillau C 1998 Messages from ultrahigh resolution crystal structures; *Curr. Opinin.. Struc. Biol.* **8** 730-737

_____, Czjzek M, Lamzin V S, Nicolas A and Cambillau C 1997 Atomic resolution (1.0Å) structure of Fusarium Solani Cutinase: Stereochemical analysis; *J Mo.l Biol* **268** 779-799

Luthy R, Bowie J U and Eisenberg D 1992 Assessment of protein models with three-dimensional profiles; *Nature* **356** 83-85

MacArthur M W and Thornton J .M 1996 Deviations from planarity of the peptide bond in peptides and proteins; *J. Mol. Biol.* **264** 1180-1195

Malakauskas S M and Mayo S L 1998 Design, structure and stability of a hyper thermophilic protein variant Nature; *Strul. Biol.* **5** 470-475

Matthews B W 1968 Solvent content of protein crystals *J. Mol. Biol.* **33** 491-497

Merklet D J, Farrington G K and Wedler F C 1981 Protein thermostability. Correlation between calculated and macroscopic parameters and growth temperature for closely related thermophilic and mesophilic bacilli; *Int. J. Pep. Protein Res.* **18** 430-432

Merritt E A 1999a Expanding the model: anisotropic displacement parameters in protein structure refinement; *Acta Crystallogr.* **D55** 1109-1117

_____ 1999b Comparing anisotropic displacement parameters in protein structures; *Acta Crystallogr.* **D55** 1997-2004

Ochi H, Hata Y, Tanaka N and Kakudo M 1983 Structure of rice ferricytochrome c at 2.0Å resolution; *J. Mol. Biol.* **166** 407-418

Oldfield T J 1992 SQUID: A program for the analysis and display of data from crystallography and molecular dynamics; *J. Mol. Graph.* **10** 247-252

Parthasarathy S and Murthy M R N 1997 Analysis of temperature factor distribution in the high-resolution protein structures; *Protein Sci.* **6** 2561-2567

Parthasarathy S and Murthy M R N 1999 On the correlation between main-cahin and side-chain atomic displacement parameters (B-values) in high resolution protein structures; *Acta Crystallogr.* **D55** 173-180

_____and _____2000a Protein Thermal Stability: insights from Atomic Displacement Parameters (B-values); *Protein Enging* **13** 9-13

_____ and _____ 2000b Correlating dynamics to conformational properties: An analysis of Atomic Displacement Parameters (B-values) in high-resolution protein structures; *Curr. Sci.* **78** 1098-1105

Petsko G A and Ringe D 1986 Fluctuations in protein structure from X-ray diffraction; *Ann. Rev. Biophys. Bioeng.* **13** 331-371

Pontius J, Richelle J and Wodak S J 1996 Deviations from standard atomic volumes as a quality measure for protein crystal structures; *J. Mol. Biol.* **264** 121-136

Querol E., Perez-Pons J A and Villarias A M 1996 Analysis of protein conformational characteristics related to thermostability; *Protein Engng* **3** 265-271

Ragone P A, Facchiano F, Facchiano A, Facciano A M, Colonna G 1981 Flexibility plot of proteins; *Protein Engng* **2** 497-504

Ramakrishnan C and Ramachandran G N 1965 Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units; *Biophys. J.* **5** 909-933

_____and _____ 1972 Stereochemical criteria for polypeptide and protein chain conformations: Effect of non-planarity and bond angle distortion at the alpha-carbon atom on the contact map for a pair of peptide units; *Int. J. Peptide Protein Res.* **4** 79-90

Rejto P A and Freer S T 1996 Protein conformational substates from X-ray crystallography; *Prog. Biophys. Molec. Biol.* **66** 167-196

Ringe D and Petsko G A 1986 Study of protein dynamics by X-ray diffraction Methods in Enzy.; **131** 389-433

Russell R J M and Taylor G L 1995 Engineering thermostability: lessons from thermophilic proteins *Curr. Opin. Biotech.* **6** 370-374

_____, Ferguson J .M C., Hough D W, Danson M J and Taylor G L 1997 The crystal structure of Citrate synthase from the hyperthermophilic archaeon pyrococcus furiosis at 1.9Å resolution; *Biochemistry* **36** 9983-9994

Sippl M J 1993 Recognition of errors in three-dimensional structures of proteins; *Proteins: Stucture, Function and Genetics* **17** 355-362

Stout C D 1993 Crystal structures of oxidised and reduced *Azotobacter vinelandii* ferridoxin at *p*H 8 and 6; *J. Biol. Chem.* **268** 25920-25927

Thompson M J and Eisenberg D 1999 Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability; *J. Mol. Biol.* **290** 595-604

Tronrud D E, Ten Eyck L F and Matthews B W 1987 An efficient general-purpose least-squares refinement program for macromolecular structures; Acta Cryst. **A43** 489-501

Usher K., De la Cruz A, Dahlquist F, Swanson R, Simon M and Remington S 1998 Crystal structures of Che Y from thermotoga maritima do not support conventional explanations for the structural basis of enhanced thermostability; *Protein Sci.* **7** 403-412

Vihinen M, Torkkila E, Riikonen P 1994 Accuracy of Protein flexibility predictions; *Proteins: structure, functions and genetics* **19** 141-149

Vogt G and Argos P 1997 Protein thermal stability: hydrogen bonding or internal packing? *Folding & Design* **2** S40-S46

_____, Woell S and Argos P 1997 Protein thermal stability, hydrogen bonds and ion pairs; *J. Mol. Biol.* **269** 631-643

Vriend, G and Sander C 1993 Quality control of protein models: Directional atomic contact analysis; *J. Appl. Cryst.* **26** 47-60

_____ 1990 WHAT I F: A molecular modeling and drug design program; *J. Mol. Graphics,* **8** 52-56

Warren G L and Petsko G A 1995 Composition analysis of (-helices in thermophilic organisms; *Protein Engng.* **9** 905-913

Wilson A J C 1949 The probability distribution of X-ray intensities; *Acta Cryst.* **2** 318-321

Wisconsin Package, Version 9.0, Genetics Computer Group (GCG), Madison, Wisconsin.

Xiao L and Honig B 1999Electrostatic contributions to the stability of hyperthermophilic proteins; *J. Mol. Biol.* **289** 1435-1444

Yip K S P, Stillman T J, Britton K L, Artymiuk P J , Baker P J, Sedelniova S E, Engel P C, Pasquo A, Chiaraluce R, Consavi V, Scandurra R and Rice D W 1995 The structure of Pyrococcus furious glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures; *Structure* **3**1147-1158