Protein Science (1997), *6*:2561–2567. Cambridge University Press. Printed in the USA.

Analysis of temperature factor distribution in high-resolution protein structures

S. PARTHASARATHY AND M.R.N. MURTHY
Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India 560 012

(RECEIVED April 15, 1997; ACCEPTED September 19, 1997)

Abstract

The temperature factors obtained from X-ray refinement of proteins at high resolution show large variations from one structure to another. However, the *B*-values expressed in units of standard deviation about their mean value (*B'*-factor) at the Cα atoms show remarkably characteristic frequency distribution. In all of the 110 proteins examined in this study, the frequency distribution exhibited a bimodal distribution. The peaks in the *B'*-factor frequency distribution occur at −1.1 and 0.4 for a bin size of 0.5. The peak at lower temperature factor corresponds largely to buried residues, whereas the peak at larger value corresponds to exposed residues. The distribution could be accurately described as a superposition of two Gaussian functions. The parameters describing the distribution are therefore characteristic of protein structures. The frequency distribution for a given amino acid over all the proteins also shows a similar bimodal distribution, although the areas under the two Gaussians differ from one amino acid to another. The area under the frequency distribution curve for any interval in *B'*-factor represents the propensity of the amino acid to occur in that interval. This propensity is related both to the hydrophilicity/hydrophobicity of the residue and the tendency of the residue to impose a different degree of rigidity on the polypeptide chain. The frequency distribution of stretches of high *B'*-factors departs appreciably from that expected for a random distribution. The correlation in the *B*-values of sequentially proximal residues is probably responsible for the bimodal distribution.

Keywords: accessibility; Gaussian functions; proteins; statistics; temperature factors

Statistical analysis has formed a large component of the research efforts put forth to understand protein structure and function, due to the enormous diversity and complexity of their structures (Johnson et al., 1994). Statistical approaches have been developed to predict the secondary structure from the primary sequence (Chou & Fasman 1974; Garnier et al., 1978; Garnier, 1990) and for testing the compatibility of model tertiary folds for a given sequence of unknown structure (Bowie et al., 1991; Luthy et al., 1991, 1992). A variety of statistical analyses has also been performed on conformational states of main chains and side chains (Dunbrack & Karplus, 1993), hydrogen bonding (Ippolito et al., 1990), water structure (Thanki et al., 1991), and topological features of secondary structural elements (Levitt & Chothia, 1976; Taylor & Thornton, 1983). Most of these analyses are concerned with protein structure and conformation. In contrast, much less attention has been paid to the atomic displacement parameter (Trueblood et al., 1996), *B*-values obtained from X-ray crystal structure analysis of proteins. The efforts in this direction have been concerned mainly with optimizing methods for prediction of antigenicity and flexibility of different polypeptide segments of a protein (Ragone et al., 1989; Karplus & Schulz 1985; Vihinen et al., 1994). In this manuscript, we examine the frequency distribution of *B*-values. The average *B*-values observed in different protein structures vary widely. However, the frequency distributions of *B*-values expressed in units of standard deviations about the mean value for the corresponding structure (*B'*-factor) are very similar. The distribution, averaged over 110 high-resolution protein structures, has maxima at −1.1 and 0.4. The distribution fits superposition of two Gaussian functions very well. We describe the frequency distribution for individual amino acids over all the proteins and for individual proteins. These distributions provide information on the relationship between amino acid residues and the rigidity imposed by them on the polypeptide chain. Analysis on short stretches of consecutive residues with high *B*-values shows that they occur less frequently than anticipated on the basis of a random distribution. This is complemented by higher than expected frequency of longer high *B*-value stretches. The observed bimodal distribution is probably related to this neighborhood effect in *B*-values.

Results and discussion

Mean profile of B'-factors

Figure 1 shows the frequency of amino acid residues in bins of *B'*-factors for bin sizes of 0.5 and 0.75. The plot corresponds to a

Reprint requests to: M.R.N. Murthy, Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India 560 012; e-mail: mrn@mbu.iisc.ernet.in.

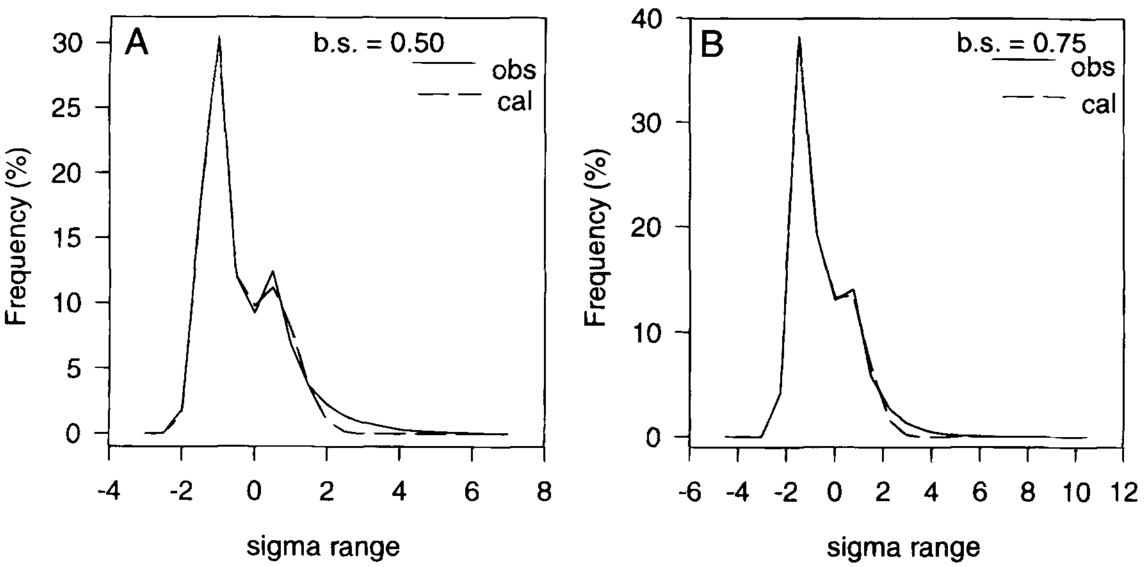

Fig. 1. Frequency distribution expressed as percentages in bins of 0.5 (**A**) and 0.75 (**B**) units of B' over 110 high-resolution protein structures. Continuous lines correspond to observed frequencies and broken lines represent the sum of the two Gaussuan functions fitted to the observed distribution.

total of 35,024 residues in the 110 selected proteins. The plot also shows the excellent fit of the analytical expression (Equation 1) to the distribution. The differences between the observed and calculated frequency for B'-factors greater than 1 is due to the low occupancy of these bins. The constants obtained, $B1 = -1.1$ and $B2 = 0.4$ represent two distinct maxima in the frequency distribution. The bimodal nature of the distribution was unaffected by the choice of the bin size. However, for bin size of 1.0 or larger, the second peak was not distinct, but appeared as a shoulder. The parameters $B1$ and $B2$ are -1.4 and 0.4 for bin size of 0.75. Because the ratio of the number of parameters (six) to the points at which the frequency distribution sampled is not high, small variations in the $B1$ and $B2$ values, depending on the sampling interval, are anticipated. However, for a given size of the bin, the parameters $B1$ and $B2$ are largely independent of the sample size of the proteins. The results, presented hereafter, correspond to a bin size of 0.5.

Figure 2A, B, C, and D shows the frequency distribution for some representative protein structures. All the proteins examined invariably exhibited bimodal distributions. Also, the constants $B1$ and $B2$ did not vary much from protein to protein. For 93% of the proteins examined, $B1$ and $B2$ values did not vary more than 10% from the mean value. These observations clearly suggest that, although the actual B-values reported for high-resolution structures of proteins are very variable, their frequency distribution expressed in the bins of B'-factors has constant features. It remains to be seen if this characteristic distribution could serve as a test of the reliability of X-ray structure refinement.

Examination of the solvent accessibility of residues falling in peak 1 in selected proteins suggested that the majority of these are inaccessible. In contrast, residues contributing to peak 2 were mostly solvent accessible. For example, in the 10 proteins for which accessibility calculations were performed, the average accessibility under peak 1 is 8.1% of the total accessibility and that under peak 2 is 27.0%. This is not surprising because the exposed residues are likely to have larger B-values. Some of the residues that belong to peak 1 but exposed may be involved in crystal contacts. However, appearance of two distinct peaks and the excellent fit of two Gaussian functions to the frequency distribution are not entirely anticipated.

Table 1. *Amino acid composition corresponding to B' range -1.25 to -0.75 and 0.25 to 0.75 in the 110 high-resolution protein structures and the overall composition in these proteins*

Amino acid	% Composition in the range		% Composition overall
	−1.25 to −0.75	0.25 to 0.75	
Ala	9.5	8.6	8.8
Cys	1.5	1.2	1.4
Asp	5.0	6.6	5.8
Glu	4.1	6.9	5.3
Phe	5.0	2.8	4.0
Gly	7.2	7.9	8.2
His	2.4	1.8	2.3
Ile	6.3	4.3	5.5
Lys	4.2	8.3	5.6
Leu	8.9	6.2	7.8
Met	2.5	1.3	2.0
Asn	4.5	5.7	4.9
Pro	3.7	5.8	4.4
Gln	3.3	4.1	3.7
Arg	4.4	4.1	4.1
Ser	6.4	7.9	7.0
Thr	6.6	7.0	6.7
Val	8.1	5.8	7.0
Trp	1.8	1.0	1.6
Tyr	4.6	2.6	3.9

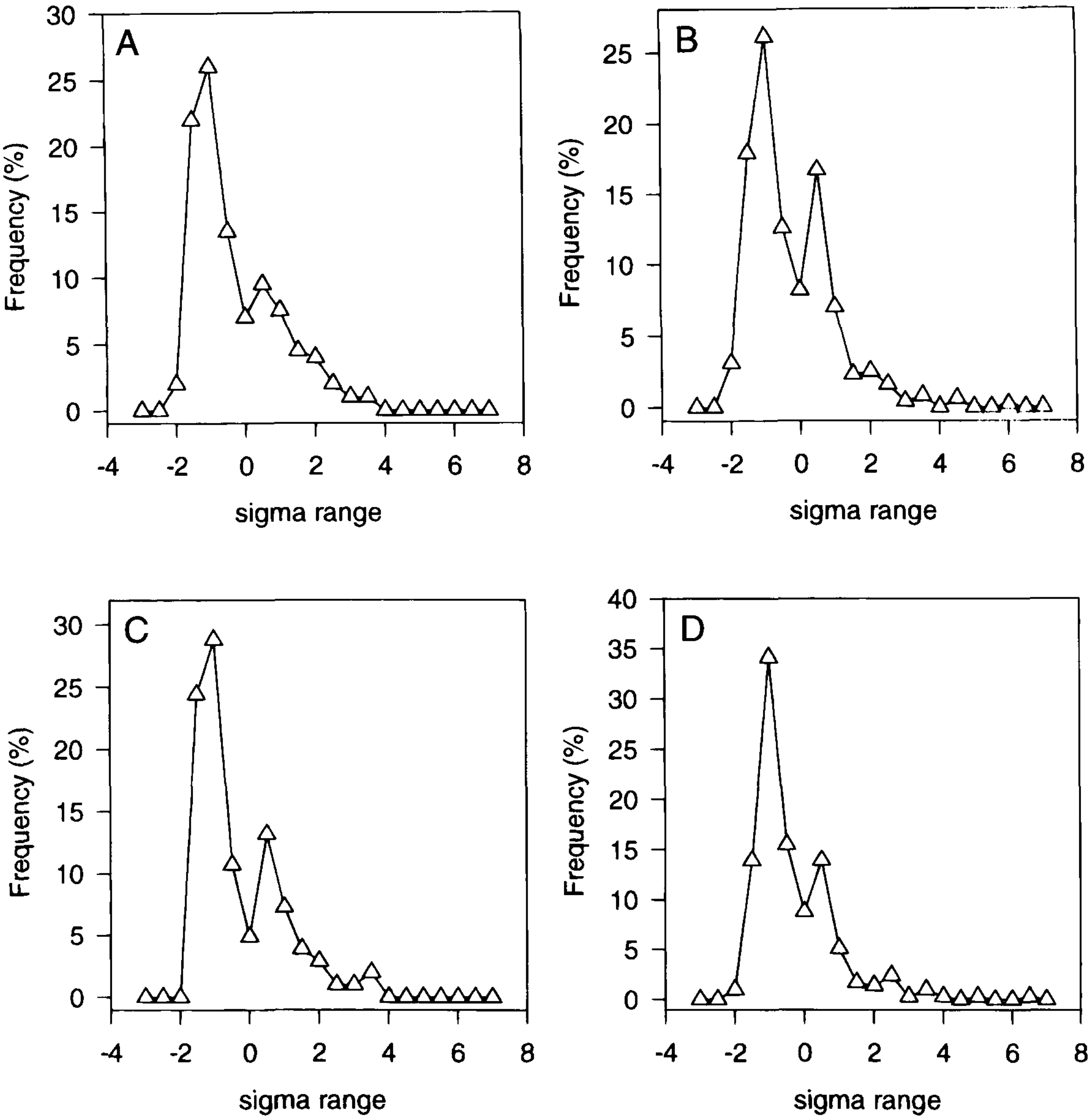

Fig. 2. Frequency distribution expressed as percentages in bins of 0.5 units of B' for some representative high-resolution protein structures. **A:** Restriction endonuclease Bam H1 (1Bam). **B:** Chitinase (1Cns). **C:** Dihydrofolate reductase (1Dyr). **D:** Ferredoxin (1Fnc).

Amino acid composition corresponding to the peaks in frequency distributions

The amino acid composition over all of the 110 proteins in the bins corresponding to the two peaks are listed in Table 1 along with the overall amino acid composition in these proteins. It is clear that the peak corresponding to the low 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.

Profiles for individual amino acid residues

Figure 3A, B, C, and D illustrates the B' frequency distribution for a number of selected residues. Most of the residues exhibit bimodal distribution, with peaks at B'-factors of -1.1 and 0.4, as observed for the overall distribution. The fit of two Gaussian functions to the distribution of every residue type was found to be good. The $B1$ and $B2$ values for different residues vary within narrow limits. (The lowest value of $B1$ is -1.211 for His, whereas the highest value of -1.009 is for Lys. Similarly, the lowest value of $B2$ is 0.217 for Trp and the highest value of 0.591 is for Ser.) The significant difference between different residues is the ratio of the areas under the two Gaussian functions. The ratio $A2/(A1 + A2)$ (p in Equation 2) represents the probability that a residue exists in the high B' region.

The differences in the probability of occurrence in the high B' region for different residues is not entirely due to the hydropathy index of the amino acid. Similarly, the differences in the amino

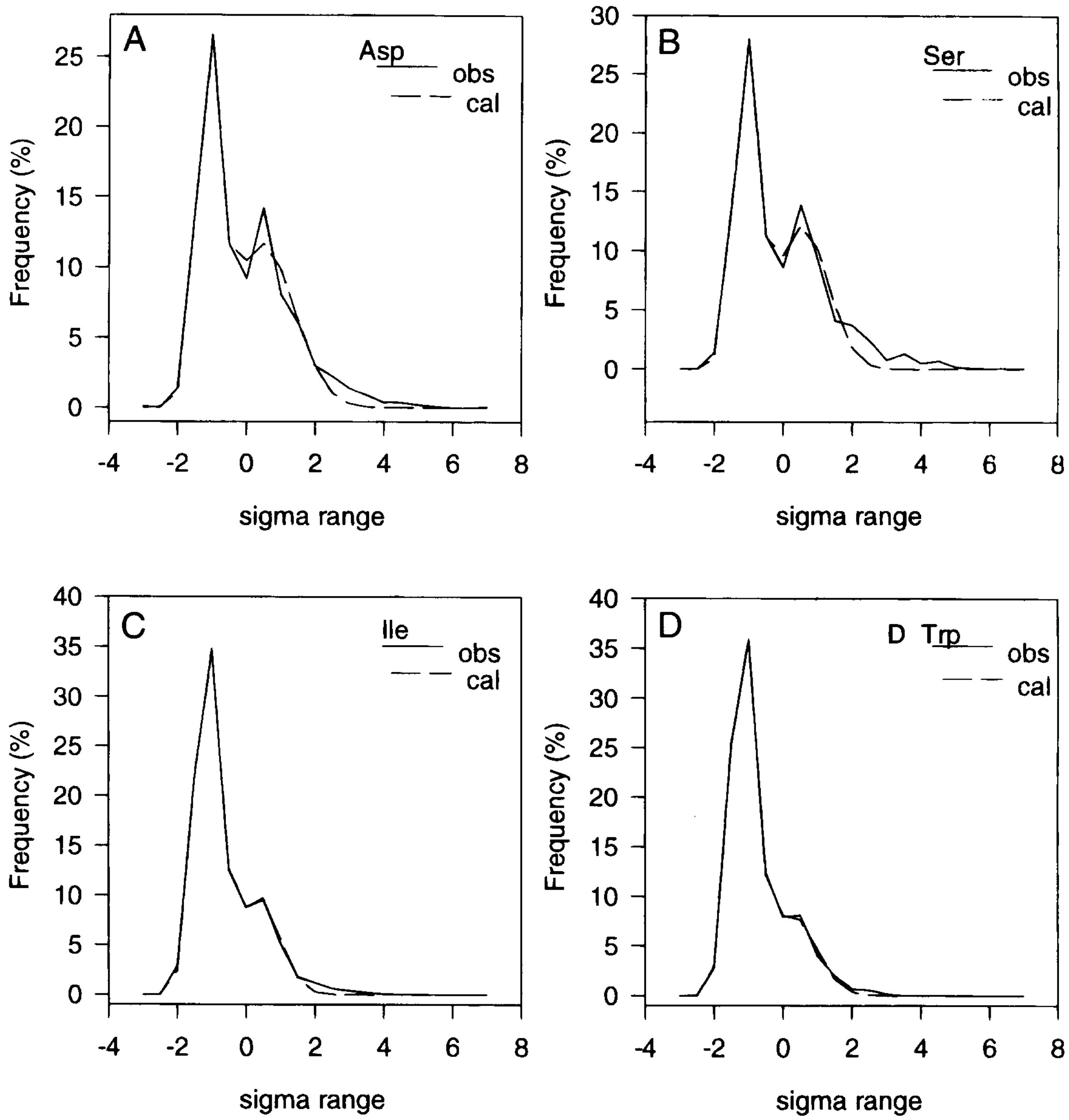

Fig. 3. Frequency distribution expressed as percentages in bins of 0.5 units of B' for representative amino acid residues. **A:** Asp, to represent charged residues. **B:** Ser, to represent small polar residues. **C:** Ile, to represent aliphatic, hydrophobic residues. **D:** Trp, to represent aromatic residues.

acid composition corresponding to peak 1 and peak 2 in the overall B' distribution (Table 1) cannot be attributed exclusively to the differences in the hydropathy properties of amino acid residues. Figure 4 illustrates the relationship between the fractional area (p values, Equation 2) and the fractional solvent accessibility parameter (f) of Janin (1979). The two parameters have a correlation coefficient of 0.52. p Values for some of the amino acids deviate appreciably from a linear relation to f. Residues that show significantly lower p value than that anticipated from the corresponding Janin parameter are mainly Met and the aromatic residues Phe, Tyr, Trp, and His. This suggests that these residues are prone to confer rigidity to the polypeptide segment. In contrast, the fraction of Gly and Pro residues that occur in regions of high B'-factors are more than their mean fractional exposure in proteins.

Profile of B-potentials along the polypeptide chain

The p values of Equation 2 represent the tendency of amino acids to occur in regions of high B'-factor. Therefore, it might be anticipated that a plot of p against the residue number in any protein might reveal segments that are likely to have large mobility. As is the usual procedure in deriving such profiles, averaging p values over a window centered on each residue helps to make the plot smoother and easy to visualize. The plot shown in Figure 5 for a thymidylate synthase mutant protein was obtained using a window of size 5. The mean p value over the residues of the window has been used for the plot. Also shown are the actually observed B-values for the same protein. These B-values have also been averaged over a window of 5. The mean correlation coefficient between the p

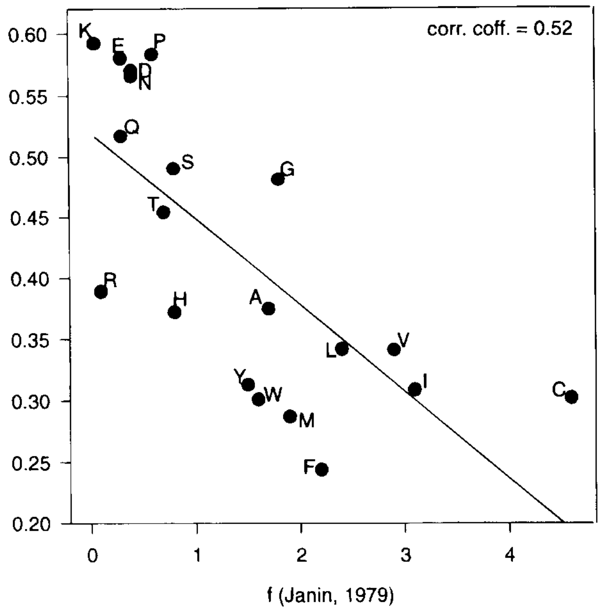

Fig. 4. Plot showing correlation between p values (Equation 2 of text) and f values of Janin (1979).

values and the B-values in the 110 proteins used for analysis is 0.34 and is as high as the value reported by other investigators using optimized window and weighting schemes (Karplus & Schulz, 1985; Vihinen et al., 1994). With these conditions, the Janin parameters based on accessibility (f), give a lower mean correlation coefficient of 0.23. The profile of p values is therefore useful in predicting flexible segments in proteins based on their amino acid sequence.

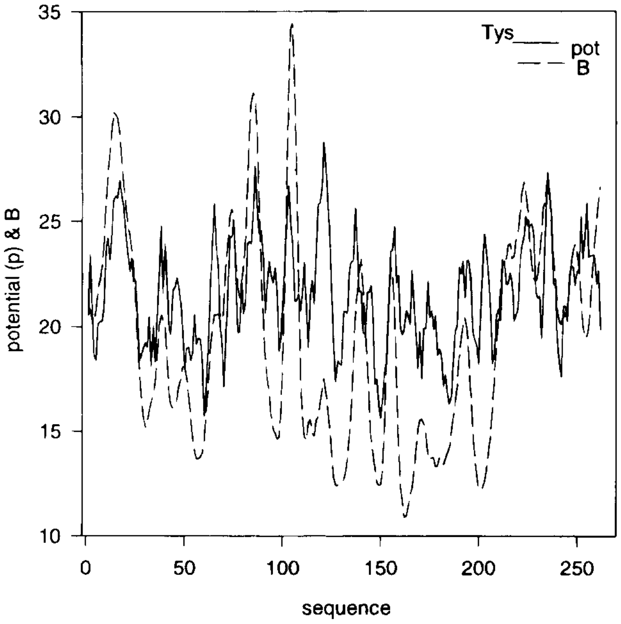

Fig. 5. Profile of p and B-values (on arbitrary scales) averaged in windows of size 5 for a mutant of thymidylate synthase (1Tys).

Table 2. *Length distribution of high B [(C_α) > ⟨B⟩ + 0.5 σ(B)] stretches in 110 high-resolution protein structures*

	Frequency	
Length	Observed	Expected[a]
1	1,066	4,646
2	436	1,008
3	235	218
4	181	47
5	125	10
6	94	2
7	76	1
8	67	0
9	51	0
10	29	0

[a] Expected frequency for random distribution of B-values.

Length distribution of occurrence of high B values

The occurrence of stretches with $B(C_\alpha) > \langle B \rangle + 0.5\ \sigma(B)$ were counted and labeled as stretches of high B-values. If f is the fraction of residues with high B-values, the probability of occurrence, given that B-values occur entirely randomly, of n consecutive high B-values will be $(1 - f)^2 f^n$. Table 2 lists the actual and observed length distribution of high B-value stretches. It might be observed that far too few single or two-consecutive resides occur with high B-values when compared to the numbers expected for a random distribution of B-values. On the other hand, stretches of three or more residues with high B-values occur more frequently than anticipated. Hence, polypeptide chains can be thought of as stretches of low B-values interrupted by high B-value stretches. It should be noted, however, that the restraints on B-values imposed during protein structure refinement might lead to biased length distributions.

The results presented in this manuscript unambiguously demonstrate that the B-values in high-resolution protein structures defined in units of standard deviation about their mean value display a bimodal distribution. The relative areas under the two peaks for different amino acids partially correlate with the tendency of amino acids to occur in the interior/exterior of the protein structure. Because there is a continuous variation in the degree of exposure of residues, the bimodal distribution cannot be explained completely in terms of the solvent accessibility properties of amino acid residues. However, the statistics of length distribution of high B stretches departs appreciably from the distribution anticipated for randomly occurring B-values. The neighborhood correlation in the B-values associated with sequentially proximal residues, as seen in the comparison of the observed and expected length distribution of high B stretches, in addition to the accessibility properties of amino acid residues, is probably the source of the bimodal nature of the B'-factor frequency distribution.

Materials and methods

Selection of high-resolution protein structures

The representative list of proteins from the Protein Data Bank (Bernstein et al., 1977) provided by Hobohm and Sanders (1994) as available during November 1996 were chosen initially for the

Table 3. *PDB codes for the 110 high-resolution structures selected for the analysis*

131L	153L	1AMP	1ARB	1ARV	1ATL
1BAM	1BP2	1CCR	1CHD	1CHMA	1CNSA
1CSEI	1CSH	1CUS	1DAAA	1DTS	1DUPA
1DYR	1EDA	1FKJ	1FNC	1GCA	1GOF
1GPR	1HMT	1HVKA	1IAE	1ISCA	1KNB
1KPTA	1LCPA	1LENA	1LFAA	1LTSA	1MOLA
1MRJ	1NAR	1NFP	1NHK	1NIF	1OVAA
1PBE	1PDA	1PGS	1PHG	1PTX	1RCF
1REC	1RGX	1RSY	1SAT	1SBP	1SNC
1SRIB	1TAG	1TAH	1TCA	1THV	1THX
1TML	1TRY	1TTBA	1TYS	1VHH	1XNB
1XYZA	2ACQ	2ALP	2AYH	2AZAA	2CBA
2CCYA	2CDV	2CPL	2CTC	2DRI	2END
2ER7E	2GSTA	2HMZA	2HTS	2MNR	2NAC
2OLVA	2PHY	2POR	2PRK	2SIL	2TGI
3CHY	3CLA	3COX	3DFR	3GRS	3PTE
3SICI	3TGL	4ENL	4FGF	4GCR	5RUBA
5RXN	5TIMA	7PCY	7RSA	8ABP	8FAB
8TLNE	9RNT				

analysis. The structures determined by NMR, those with more than 25% sequence identity and those determined with resolution less than 2.0 Å or R-factor greater than 0.2, were rejected from this list. Of the remaining 164 structures, 110 protein coordinates were complete in all respects and were found to be acceptable. The PDB codes of the selected proteins are listed in Table 3.

Frequency distribution of B-factors

For each selected protein, the following quantities were computed. The mean of B at C_α positions, $\langle B\rangle = \Sigma B_i/N$, where B_i is the B-value associated with C_α of the ith residue and N is the total number of residues in the protein; $\sigma^2(B) = \Sigma(B - \langle B\rangle)^2/N$; modified B (B'-factor) representing normal variate as $B' = (B - \langle B\rangle)/\sigma(B)$.

Residues were divided into 0.5-unit ranges in B' and the frequency of occurrence of residues in each bin was counted. If f_{ijk} is the number of times a residue of type i occurs in protein j in the kth bin of B', then $\Sigma_{jk}f_{ijk}$ is the total number of times the residue i occurs in the structures selected for analysis, $\Sigma_{ij}f_{ijk}$ is the number of residues in the kth bin of B' in all the structures selected for analysis, $\Sigma_i f_{ijk}$ is the number of residues that occur in kth bin in protein j, $\Sigma_j f_{ijk}$ is the number of occurrence of amino acid i in bin k over all the proteins. Frequency distribution for individual amino acids over all the proteins ($\Sigma_j f_{ijk}$), individual proteins over all the amino acids ($\Sigma_i f_{ijk}$), and the distribution including all proteins and all amino acids ($\Sigma_{ij}f_{ijk}$) were fitted to the equation

$$f = k1e^{-k2*(B'-B1)^2} + k3e^{-k4*(B'-B2)^2}, \qquad (1)$$

where, $k1$, $k2$, $k3$, $k4$, $B1$, and $B2$ are parameters. These parameters were refined by a least-squares procedure. The areas under the Gaussian functions are given by

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

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

The fractional area under the second peak for each amino acid was computed as

$$p = A2/(A1 + A2). \qquad (2)$$

These fractional areas (p) represent the probability or propensity that the given amino acid occurs with high B' value.

Surface accessibility calculations

Surface accessibility computations were performed based on Lee and Richards (1971) algorithm. The calculation was performed on isolated protein monomers. Hence, the results presented here do not take into account the effect of lattice or subunit contacts. The atomic accessibility values computed were appropriately summed to provide residue accessible surface areas. These areas were divided by the maximum accessible area for the particular residue in Gly-X-Gly sequence in a completely extended conformation (Creighton, 1992).

Profile of B′ potentials

A sliding window averaging technique was used for calculating the profile of B' potentials along the polypeptide sequence. A window of 5 was selected. The average of the B' potential of the five residues in the window was found and associated with the central residue. The resulting profile was evaluated for its agreement with observed B-values. No optimization of the window size or weighting scheme for the residues in the window was attempted.

Acknowledgments

We thank DST (M.R.N.M.) and CSIR (M.R.N.M., S.P.) India for financial support. We thank Dr. S. Baranidharan for useful discussions.

References

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, 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.

Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:222–245.

Creighton TE. 1992. *Proteins: Structures and molecular properties*. New York: Freeman and Company. pp 142–143.

Dunbrack RL Jr, Karplus M. 1993. Backbone-dependent rotamer library for proteins. *J Mol Biol 230*:543–574.

Garnier J. 1990. Protein structure prediction. *Biochimie 72*:513–524.

Garnier J, Osguthrope DJ, Robson B. 1978. Analysis of the accuracy and implications of simple method for predicting the secondary structure of globular proteins. *J Mol Biol 120*:97–120.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Ippolito JA, Alexander RS, Christianson DW. 1990. Hydrogen bonding stereochemistry in protein structure and function. *J Mol Biol 215*:457–471.

Janin J. 1979. Surface and inside volumes in globular proteins. *Nature 277*:491–492.

Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL. 1994. Knowledge-based protein modeling. *CRC Rev Biochem Mol Biol 29*:1–68.

Karplus PA, Schulz GE. 1985. Prediction of chain flexibility in proteins. *Naturwissenschaften 72*:212–213.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379–400.

Levitt M, Chothia C. 1976. Structural patterns in globular proteins. *Nature 261*:552–558.

Luthy R, McLachlan AD, Eisenberg D. 1991. Secondary structure-based profiles: Use of structure conserving scoring tables in searching protein sequence data bases for structural similarities. *Proteins Struct Funct Genet 10*:229–239.

Luthy R, Bowie JU, Eisenberg D. 1992. Assessment of protein models with three-dimensional profiles. *Nature 356*:83–85.

Ragone PA, Facchiano F, Facchiano A, Facchiano AM, Colonna G. 1989. Flexibility plot of proteins. *Protein Eng 2*:497–504.

Vihinen M, Torkkila E, Riikonen P. 1994. Accuracy of protein flexibility predictions. *Proteins Struct Funct Genet 19*:141–149.

Taylor WR, Thornton JM. 1983. Prediction of super-secondary structure in proteins. *Nature 301*:540–542.

Thanki N, Umrania Y, Thronton JM, Goodfellow JM. 1991. Analysis of protein main-chain solvation as a function of secondary structure. *J Mol Biol 221*:669–691.

Trueblood KN, Burgi HB, Burzlaff H, Dunitz JD, Gramaccioli CM, Schulz HH, Shmueli U, Abrahams SC. 1996. Atomic displacement parameters nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. *Acta Crystallogr A 52*:770–781.