J. Appl. Cryst. (1996). **29**, 100–104

Knowledge-Based *B*-Factor Restraints for the Refinement of Proteins

DALE E. TRONRUD

Howard Hughes Medical Institute and Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA. E-mail: dale@uoxray.uoregon.edu

(Received 10 May 1995; accepted 13 October 1995)

Abstract

A new type of restraint for the *B* factors of atoms in low-to-moderate resolution models of proteins is proposed. This restraint incorporates the knowledge that the *B* factors of two atoms bonded to each other may be systemically different. In addition, some bonded pairs of atoms will be more consistent from structure to structure than others. With the use of *B*-factor restraints of this type, it is possible to construct models whose *B* factors are consistent with accurately known protein structures, even though only low-resolution crystallographic data are available.

1. Introduction

The principal problem with macromolecular refinement is the lack of diffraction data in relation to the number of parameters in the model. In many cases, the diffraction from a crystal cannot be measured beyond 3 Å or so because the intensities become too weak. In a typical case, with the solvent content assumed to be 50%, there will be only four to five reflections per ordered atom. The standard form of the structural models used in refinement contains four parameters per atom. The diffraction data set does not contain sufficient information to define all the parameters of the model.

To supplement the diffraction data, one generally adds as observations relationships between the atoms of the model derived from higher resolution models. In the restrained refinement method currently in general use [*PROLSQ* (Hendrickson & Konnert, 1980), *XPLOR* (Brünger, Kuriyan & Karplus, 1987) and *TNT* (Tronrud, Ten Eyck & Matthews, 1987)], the positions of the atoms are restrained to conform to the expected bond lengths, bond angles and other positional information. Considerable attention has been paid to the derivation of precise libraries of stereochemical restraints using small-molecule models (Kennard, 1968; Levitt, 1974; Vijayan, 1976; Kennard & Taylor, 1982; Saenger, 1983; Allen *et al.*, 1987; Engh & Huber, 1991). However, very little has been done to identify restraints that could be applied to the *B* factors of a macromolecular model. This paper describes an attempt to restrain *B* factors using a library of restraints derived from a small collection of well refined protein models. The proposed method is compared to the method currently in use.

2. Existing methods and limitations

When a model of a macromolecule is refined against low-to-moderate-resolution diffraction data (~2 Å or lower), the *B* factors of atoms fluctuate wildly from one atom to the next. The lower the resolution, the larger the fluctuations. Since *B* factors are usually interpreted as a measure of the amount of motion that that atom experiences, it is not plausible for atoms bonded to one another to exhibit substantially different *B* factors.

The simplest way to remove this problem is to add a restraint to the refinement function that causes the *B* factors of atoms bonded to each other to be similar. The *B*-factor restraints in *PROLSQ* (Hendrickson & Konnert, 1980) and *XPLOR* (Brünter, Kuriyan & Karplus, 1987) are based upon the work of Konnert & Hendrickson (1980). In this approach, it is assumed that in each pair of bonded atoms one of the atoms is 'riding' on the other. Its *B* factor will contain the motion of the first atom as well as its own relative motion. Consider three atoms, *A*, *B* and *C*, with *A* bonded to *B* and *B* in turn bonded to *C*. Since *A* is directly coupled to *B*, the relative motion along the bond is very restricted. One can derive a restraint on the component of the anisotropic *B* factor along the bond direction. Konnert & Hendrickson (1980) recommend that this restraint be enforced within a root-mean-square value of 2.5 Å^2. Konnert & Hendrickson also describe a restraint on the motion, and therefore the anisotropic *B*-factor component, along the line connecting two atoms at the ends of a bond angle, *i.e.* between *A* and *C*. This restraint was given a target value of 8 Å^2.

Konnert & Hendrickson briefly discuss the application of restraints on isotropic *B* factors. Since an isotropic *B* factor has no directional variation, one simply minimized the difference in *B* factor for each bonded pair of atoms. No target value was suggested for the isotropic restraint.

In order to illustrate the problem posed by the choice of an isotropic restraint target value, consider the case of the Oγ atom in a serine residue. This atom is bonded to the Cβ atom and the anisotropic components of the two atoms along this bond should be tied together with a precision of 2.5 Å^2. The Cα—Cβ—Oγ bond angle also restricts the motion of the Oγ atom. To reflect this, the anisotropic *B* factor of Oγ along the Cα—Oγ direction can be restrained with a precision of 8 Å^2. The remaining anisotropic component, perpendicular to

ISSN 0021-8898 © 1996

the two mentioned, is essentially unrestricted, because the atom can rotate about the N—Cα—Cβ—Oγ torsion angle.

Since the isotropic B factor is the mean of the three principal components of the anisotropic B factor, it should be restricted less than the most restricted component and more than the least restricted component. The median of the three restraints, 8, is a good estimate of the target value for the isotropic B-factor restraint.

PROLSQ and *XPLOR* apply restraints on the isotropic B factors; however, they inappropriately use the anisotropic B-factor target. In these programs, the target value for the root-mean-square difference in B factor for bonded atoms is very small, typically in the neighborhood of 2 Å^2. The models produced by these programs underestimate the variability in B factor from atom to atom.

This problem was first noted by Yu, Karplus & Hendrickson (1985). The authors compared the fluctuations of the molecular dynamic simulation of a small protein with the B factors determined crystallographically restrained by this method. The atoms in the simulation showed variations in amplitude of motion two to three times larger than those allowed by the B-factor restraints.

A more significant problem with these restraints is the lack of provision for the expectation that certain atoms move more than the atoms to which they are bonded. In the example above, it is expected that an Oγ atom will have a B factor that is larger than that of the Cβ atom. The restraint, however, attempts to equalize the B factors. This will cause a bias in the model such that B factors of serine Oγ atoms will be systematically underestimated.

In addition, one would expect certain pairs of atoms to exhibit greater consistency in their B factors than others. One would like an individual standard deviation for each class of pairs. For instance, one would expect that the B factors of the backbone Cα and N atoms would show roughly the same correspondence throughout the structure. In contrast, some lysine side chains are well ordered while others are disordered. In the former case, the atoms will typically display only modest increases in B factors as one moves away from the main chain. In the disordered cases, however, the B factors will increase rapidly. Restraints derived from averaging of the lysine side chains must clearly be given a smaller weight than those applied to the main chain atoms.

While the method of Konnert & Hendrickson (1980) does decrease the fluctuations in B factor, it does not allow the variations in B factors that are to be expected on the basis of current knowledge of macromolecular structures.

3. The proposed method

It is proposed that the B factors of macromolecular models be restrained by

$$f = \sum_{i}^{\text{all bonds}} [1/\sigma(i)^2][\Delta_i - (B_1 - B_2)]^2, \qquad (1)$$

where B_1 and B_2 are the current B factors of the two bonded atoms, 1 and 2. Δ_i is the expected increase in B factor when moving from atom 1 to atom 2. $\sigma(i)$ is a measure of the confidence of Δ_i. The parameters of the model are adjusted to minimize this function, while simultaneously minimizing the residual in the diffraction data restraints.

One cannot have a separate Δ_i for each bond in the molecule; one would not know what value to use in each instance. The bonds must be categorized in some fashion to reduce the number of standard values required. In this work, a systematic study was not performed to determine an optimal parametrization. Instead, the method used in the *TNT* refinement package for the description of stereochemistry was chosen. In this scheme, all the bonds in the main chain are considered to be in the same class, while the bonds in the side chains are classified by the amino-acid type.

4. Determination of standard values

Before these new restraints can be applied to a model, the values of Δ_i and $\sigma(i)$ must be determined from accurately known structures in which we have confidence.

The choice of the basis set is complicated because on the one hand one would like these models to be based on very high resolution diffraction data sets, but on the other hand the motion of the atoms in the models should be similar to those in 'typical' macromolecules. In particular, one would not expect the motion seen in crystals of small molecules to be representative of the motions of macromolecules.

The values of Δ_i and $\sigma(i)$ must therefore be derived from protein molecules. The resolution of the diffraction data must be fairly high, *e.g.* at least 1.7 Å. The B factors of the models should not have been restrained by any factors, such as those that one is attempting to derive. This restriction prevents any bias caused by either the traditional or new restraints. Because the exclusion of low-resolution diffraction terms causes systematic errors in the B factors, the models used to derive the standard values must have been refined against all of the data. Since most models produced by refinement with *PROLSQ* or *XPLOR* are routinely subjected to B-factor restraints, they cannot be used to determine standard values for the new restraining function. This excludes a very large proportion of the structures in the Protein Data Bank. Therefore, it was necessary to resort to a set of four structures, with a total of about 900 amino-acid residues. The structures are listed in Table 1. The library of standard values was

Table 1. *Reference models*

These are the models used to derive values for Δ_i and $\sigma(i)$. The T4 lysozyme model used was that of the mutant form (C54T, N68C, A93C, C97A) because the quality of the diffraction data of that mutant was judged to be superior to all the others. *R* is the resolution.

Protein	*R* (Å)	PDB No.	Reference
Thermolysin	1.6	8TLN	Holland *et al.* (1982)
Goose lysozyme	1.6	153L*	Weaver, Grütter & Matthews (1995)
Mutant T4 lysozyme	1.7	139L†	Heinz & Matthews (1994)
γ Chymotrypsin	1.6	1GCT*	Dixon & Matthews (1989)

* Model subjected to additional refinement. Modifications include the rotation of some side chains and the re-evaluation of each solvent molecule, as well as additional cycles of refinement. † Very minor changes were made when the model was deposited. The analysis in this paper was performed on the original version.

generated by calculation of the mean *B*-factor change for each type of bond in all the models, as well its standard deviation.

Parameters for nucleic acids have not been defined due to the lack of a basis set.

A benefit of this survey is that it provides confidence limits for the *B* factors derived by the refinement procedure. If the *B* factors for these models were unreliable all of the $\sigma(i)$ parameters would be quite large. In fact, the mean $\sigma(i)$ for bonds between atoms that are well ordered (*e.g.* main-chain atoms and the side-chain atoms of hydrophobic side chains) is about 5 Å^2. This implies that the 95% confidence interval for the *B* factors of these models is approximately ±9.8 Å^2 [= $1.96\sigma(i)$ because the integral of the normal distribution from -1.96 to 1.96 is 0.95].

5. Results

Table 2 summarizes the analysis of the *B* factors for the peptide group, the 20 amino acids and the disulfide bond. As expected, adjacent atoms within peptide groups have, on average, fairly similar *B* factors. In contrast, a number of the amino acids display substantial increases in their *B* factors as one proceeds from the α-carbon toward the more distal atoms.

As a specific example, consider threonine and valine. These amino acids are isostructural but have very different hydrophobicity and occur in very different chemical environments. In threonine, the *B* factor of the Cβ atom is 8 Å^2 larger than that of Cα, while in valine the corresponding difference is only 4 Å^2. This observation is consistent with valine's preference for being located in the well ordered core of the protein. In addition, the standard deviations, $\sigma(i)$, are larger for the restraints in threonine because there is more variability in the location of these side chains. Threonine residues are sometimes buried (and ordered) and at other times solvent-exposed (and disordered), while valine residues are almost always internal (and well ordered). The smaller standard deviation for valine allows the *B*-factor correlation restraint to be held more tightly in this case, because one has greater confidence in the validity of the restraint.

In addition to the library itself, one must determine how closely new models will be forced to agree with these restraints. In other words, what is the target value for the final root-mean-square deviation from the ideal values. The target value can be estimated using cross validation. In cross validation the statistical analysis is repeated with a subset of data excluded and the agreement between the subset and the test analysis is monitored. In this case, a new library of Δ_i and $\sigma(i)$ restraints is determined with one protein excluded and the root-mean-square deviation of the bonds in the excluded protein compared to the test library is calculated. This calculation is repeated with each protein excluded in turn, resulting in 'complete' cross validation. The results of this test are listed in Table 3. In refinement protein models are expected to agree with this library with a root-mean-square error of 6.5 Å^2.

6. Example refinement

To demonstrate the effect of the proposed restraints, two refinements were performed with the same starting coordinate file. The structure was a complex of thermolysin with the inhibitor phosphoramidon (Weaver, Kester & Matthews, 1977). Even though the crystals diffracted well, the data were collected to only 2.3 Å resolution. The starting model was constructed by placing the unrefined inhibitor model (Weaver *et al.*, 1977) into the model of inhibitor-free thermolysin described previously (Holmes & Matthews, 1982). The thermolysin model had been refined with *PROLSQ* (Hendrickson & Konnert, 1980) and the *B* factors restrained in the usual fashion for that program.

In the first refinement, this model was refined for 30 cycles using the *TNT* refinement package (Tronrud, Ten Eyck & Matthews, 1987) without the application of any stereochemical restraints on the *B* factors. The final model had an overall *B*-factor discrepancy (based on the new library) of 14.0 Å^2. In the second refinement, the new library was enforced and the final model had a root-mean-square *B*-factor discrepancy of 6.4 Å^2.

Even though the agreement to the restraints was much improved, the *R* value rose only 0.1% (from 13.6 to 13.7%). This result tends to confirm that the restraints add information that is not present in diffraction data of 2.3 Å resolution.

7. Summary

A new formulation is described for restraining the *B* factors of low- to moderate-resolution protein models based on stereochemistry. The procedure can be implemented simply and requires very little additional computation.

Table 2. *Temperature-factor restraint library*

The sign of Δ_i depends on the order of consideration of the two atoms. The convention chosen here is that the *B* factor of the first atom is subtracted from that of the second. Therefore, a Δ_i of 8.2 Å^2 for CA and CB of threonine means that, on the average, the *B* factor of CB is 8.2 Å^2 larger than that of the corresponding CA. One aspartic acid residue in the sample does not contain atoms beyond CB. This absence results in a decrease in the sample size for the residue's last three bonds.

Amino acid	Atom 1	Atom 2	Δ_i (Å^2)	$\sigma(i)$ (Å^2)	Sample
Peptide	N	CA	−0.2	4.6	898
	CA	C	3.7	7.4	
	C	+N	−3.5	6.6	
	C	O	−0.5	6.3	
ALA	CA	CB	0.7	3.5	80
ARG	CA	CB	0.3	8.5	36
	CB	CG	10.6	16.5	
	CG	CD	10.1	21.3	
	CD	NE	8.8	22.7	
	NE	CZ	14.4	21.3	
	CZ	NH1	−12.5	23.5	
	CZ	NH2	−13.1	23.4	
ASN	CA	CB	0.9	10.8	51
	CB	CG	23.1	25.0	50
	CG	OD1	−5.6	21.1	50
	CG	ND2	−8.1	18.9	50
ASP	CA	CB	1.8	4.5	55
	CB	CG	9.8	14.3	
	CG	OD1	0.5	9.7	
	CG	OD2	5.8	14.3	
CYS	CA	CB	0.3	3.8	16
	CB	SG	3.9	3.2	
GLN	CA	CB	1.4	4.1	37
	CB	CG	15.7	22.8	
	CG	CD	19.7	24.1	
	CD	OE1	−4.3	18.4	
	CD	NE2	−7.7	22.3	
GLU	CA	CB	1.4	3.5	27
	CB	CG	10.5	13.5	
	CG	CD	20.0	25.1	
	CD	OE1	−5.9	33.9	
	CD	OE2	−6.1	26.9	
HIS	CA	CB	0.8	3.8	16
	CB	CG	3.4	3.8	
	CG	ND1	1.7	4.0	
	CG	CD2	1.7	5.4	
	ND1	CE1	−0.6	4.3	
	CE1	NE2	0.3	3.3	
	CD2	NE2	−0.3	3.7	
ILE	CA	CB	3.1	4.2	52
	CB	CG1	0.4	5.0	
	CG1	CD1	7.4	17.2	
	CB	CG2	0.7	6.9	
LEU	CA	CB	0.1	3.6	56
	CB	CG	4.7	5.3	
	CG	CD1	0.2	5.6	
	CG	CD2	2.5	9.1	
LYS	CA	CB	2.6	5.8	57
	CB	CG	14.8	21.2	
	CG	CD	10.3	18.9	
	CD	CE	11.2	24.1	
	CE	NZ	9.3	24.2	
MET	CA	CB	2.3	3.4	12
	CB	CG	3.7	4.3	
	CG	SD	4.2	4.7	
	SD	CE	−3.3	4.5	
PHE	CA	CB	1.8	3.8	24
	CB	CG	1.6	4.5	
	CG	CD1	3.1	6.2	
	CG	CD2	2.2	6.2	
	CD1	CE1	0.9	5.3	
	CD2	CE2	2.8	5.3	
	CE1	CZ	−0.9	6.7	
	CE2	CZ	−1.9	5.1	
PRO	CA	CB	1.8	3.9	25
	CB	CG	4.8	5.8	
	CG	CD	−4.4	4.9	
	CD	N	−0.1	5.6	
SER	CA	CB	3.6	7.1	68
	CB	OG	8.8	13.9	
THR	CA	CB	8.1	13.6	72
	CB	OG1	−0.4	14.7	
	CB	CG2	−1.4	12.6	
TRP	CA	CB	−1.3	3.8	17
	CB	CG	0.9	3.5	
	CG	CD1	2.9	2.8	
	CD1	NE1	−0.6	4.2	
	NE1	CE2	1.4	4.3	
	CE2	CZ2	−0.7	4.2	
	CZ2	CH2	0.1	6.0	
	CZ3	CH2	0.6	3.0	
	CE3	CZ3	1.5	3.8	
	CD2	CE3	1.3	3.4	
	CG	CD2	−0.2	3.0	
	CD2	CE2	3.9	5.3	
TYR	CA	CB	1.2	4.7	48
	CB	CG	1.7	6.0	
	CG	CD1	2.0	4.8	
	CG	CD2	0.8	4.6	
	CD1	CE1	0.8	7.9	
	CD2	CE2	0.9	4.4	
	CE1	CZ	6.3	14.6	
	CE2	CZ	7.4	12.2	
	CZ	OH	0.9	13.1	
VAL	CA	CB	3.9	4.8	66
	CB	CG1	−0.2	4.3	
	CB	CG2	−0.2	3.8	
SS	SG	+SG	0.0	2.1	9

It permits *B* factors to have the variability observed in reliably determined structures but damps the fluctuations caused by the limited diffraction data.

The procedure is intended for cases where the diffraction data are limited to a resolution of about 2 Å or less. It could be used at higher resolution, although the

Table 3. *Estimation of target value*

One can estimate the level to which these restraints should be enforced with cross validation. One at a time, each protein is removed from the basis set and a library of restraints is derived from the remaining proteins. The direct validation is the root-mean-square error found in the proteins used to generate the library. The cross validation is the root-mean-square error of the bonds in the protein excluded. One expects that new protein models should be consistent with the restraint library to the average value of the cross validation. The target value for this restraint is 6.5 Å^2.

Protein excluded	Direct validation (Å^2)	Cross validation (Å^2)
Thermolysin	5.453	6.055
Goose lysozyme	5.459	6.368
Mutant T4 lysozyme	5.935	6.186
γ Chymotrypsin	5.128	7.366
Average (standard deviation)	5.49 (33)	6.49 (60)

present restraint library is limited by the resolution of the reference protein structures (1.6–1.7 Å) and the small sample size.

The restraints and the overall procedure are currently available in the *TNT* refinement package (Tronrud, Ten Eyck & Matthews, 1987) for the refinement of proteins. A further survey including nucleic acids and common enzyme co-factors still needs to be done.

This work was supported in part by NIH grant GM20066 to B. W. Matthews.

References

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orphen, A. G. & Taylor, R. (1987). *J. Chem. Soc. Perkin Trans.* **2**, S1–S19.

Brünger, A., Kuriyan, K. & Karplus, M. (1987). *Science*, **235**, 458–460.

Dixon, M. M. & Matthews, B. W. (1989). *Biochemistry*, **28**, 7033–7038.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Heinz, D. W. & Matthews, B. W. (1994). *Prot. Eng.* **7**, 301–307.

Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, ch. 13, pp. 13.01–13.26. Bangalore: Indian Academy of Sciences.

Holland, D. R., Tronrud, D. E., Pleyk, H. W., Flaherty, M., Stark, W., Jansonius, J. N., McKay, D. B. & Matthews, B. W. (1992). *Biochemistry*, **31**, 11310–11316.

Holmes, M. A. & Matthews, B. W. (1982). *J. Mol. Biol.* **160**, 623–639.

Kennard, O. (1968). *International Tables for X-ray Crystallography*, Vol. III, pp. 275–276. Birmingham: Kynoch Press. (Present distributor Kluwer Academic Publishers, Dordrecht.)

Kennard, O. & Taylor, R. (1982). *J. Am. Soc. Chem.* **104**, 3209–3212.

Konnert, J. H. & Hendrickson, W. A. (1980). *Acta Cryst.* A**36**, 344–350.

Levitt, M. (1974). *J. Mol. Biol.* **82**, 393–420.

Saenger, W. (1983). *Principles of Nucleic Acid Structure*, pp. 70 and 86. New York: Springer-Verlag.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.

Vijayan, M. (1976). *CRC Handbook of Biochemistry and Molecular Biology*, 3rd ed., *Proteins*, Vol. III, pp. 742–759. Cleveland: Chemical Rubber Company.

Weaver, L. H., Grütter, M. G. & Matthews, B. W. (1995). *J. Mol. Biol.* **245**, 54–68.

Weaver, L. H., Kester, W. R. & Matthews, B. W. (1977). *J. Mol. Biol.* **114**, 119–132.

Yu, H.-A., Karplus, M. & Hendrickson, W. A. (1985). *Acta Cryst.* B**41**, 191–201.