Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures

Manfred S. Weiss

EMBL Hamburg Outstation, c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany

Correspondence e-mail: msweiss@embl-hamburg.de

Received 21 July 2007
Accepted 21 October 2007

Macromolecular models refined against X-ray diffraction data are typically described by a set of atomic coordinates and atomic displacement parameters (ADPs). Although it is intuitively obvious that the two cannot be independent of each other and although over time many attempts have been made to relate them to each other, such approaches have so far not been utilized in macromolecular structure refinement. It is demonstrated here that up to 50% of the total ADP variation in macromolecular structures may be successfully predicted solely based on the atomic coordinates and just three additional parameters per structure. This finding may have serious implications in macromolecular structure refinement, particularly at low resolution, as well as in structure validation.

1. Introduction

A biological macromolecule can only carry out its function because it possesses a certain three-dimensional structure or because it can adapt its shape according to external stimuli, such as the presence or absence of certain interaction partners. A macromolecular model is typically described as a set of spatial coordinates for each atom making up the molecule and a corresponding set of atomic displacement parameters (ADPs) or atomic temperature factors (commonly referred to as *B* factors). Whereas the coordinates describe the three-dimensional structure as such or the architecture of the molecule, the ADPs provide information about the flexibility of the structure. It is often the latter aspect, the structural flexibility, without which it is simply impossible to understand and rationalize the function of a molecule.

At present, more than 40 000 macromolecular structures are available from the Protein Data Bank (Berman *et al.*, 2000), of which about 85% are the outcome of a crystallographic experiment. Whether the crystallographic structure determination proceeded *via* molecular replacement or *via* experimental phasing, the last step of a structure determination is always the refinement of the structure against the experimentally derived X-ray intensities or structure-factor amplitudes. In this refinement step, the spatial coordinates of each atom and its corresponding ADP values (one parameter if isotropic *B*-factor refinement is performed, which is the case for most of the structures in the PDB) are typically considered independent variables which are also refined independently.

From the mechanistic point of view, however, the flexibility of a given atom must not be considered independent of its surrounding atoms. Two atoms within the same environment will always be able to move about their equilibrium position in

a similar fashion. Consequently, structure and flexibility are dependent on each other and it should therefore be possible to interrelate the architecture of a molecule, *i.e.* its three-dimensional coordinates, to the atomic flexibility, *i.e.* its ADP values.

Over time, many methods have been investigated to achieve this task. The simplest approach is the translation–libration–screw (TLS) model (Schomaker & Trueblood, 1968), which views the crystallized protein as a rigid body or as an assembly of rigid bodies undergoing various motions inside the crystalline environment. Using the TLS model, ADP values can be assigned to each individual atom of a given rigid body as a function of its position relative to the origin for the TLS tensor. Another approach to identify rigid groups in macromolecular structures originates from the application of graph theoretical approaches to a bond network describing the covalent and noncovalent forces holding the molecule together (Jacobs *et al.*, 2001; Gohlke *et al.*, 2004). ADP values have also been correlated with atomic fluctuations in molecular-dynamics (MD) simulations with potential functions of varying complexity (MacKerell *et al.*, 1998; Hinsen & Kneller, 1999; Higo & Umeyama, 1997) or have been predicted based on normal-mode analyses of protein structures (Levitt *et al.*, 1985; Tirion, 1996; ben-Avraham & Tirion, 1998). More recently, Gaussian network models (GNMs) have been used for this purpose (Bahar *et al.*, 1997, 1998; Haliloglu *et al.*, 1997; Haliloglu & Bahar, 1999; Kundu *et al.*, 2002). GNMs describe a macromolecule as a collection of atoms connected by springs, where the atoms fluctuate about their mean positions, and can be set up with the individual molecule either outside or inside its crystalline environment. Despite the fact that these approaches are computationally rather expensive, their success has been modest. A simple extension of GNMs has been used by Halle (2002), who assumed, based on some crude approximations, an inverse relationship between the ADP values of a macromolecular structure and the local packing density of the atoms of the structure. The work presented here builds on the ideas of Halle (2002) and relates refined ADP values to various functions containing the local packing density of atoms of macromolecular structures inside their crystalline environment. It can be shown that very simple three-parameter models are sufficient to predict the ADP variation in a protein structure, which on average accounts for 50% of the improvement in the crystallographic R factor compared with using just one average ADP value for all atoms in the structure refinement.

2. Methods

2.1. The database

A nonredundant list of protein structures was prepared from the Protein Data Bank (Berman *et al.*, 2000) on 26 March 2007 using the PISCES server (http://dunbrack.fccc.edu/PISCES.php; Wang & Dunbrack, 2003). The input thresholds for culling the PDB were as follows: maximum 25% sequence identity, maximum resolution between 1.5 and 1.8 Å, maximum R factor 20%, sequence length 100–999 amino-acid residues. In addition, non-X-ray structures and C^{α}-only models were excluded. By using this relatively narrow resolution range, it was ensured that the ADP values considered are equally reliable (Carugo & Argos, 1999) and that they do not contain any systematic differences caused by differing resolutions.

2.2. Calculation of the contact numbers

The PDB entries were downloaded from the local EMBL Hamburg copy of the PDB provided they had an associated structure-factor file. The structures were then placed in their crystallographic environment using the program *CPC* (O. Carugo, University of Pavia, Italy, personal communication). For this purpose, all water and ligand molecules associated with the protein structure were disregarded. All crystallographic neighbours were generated when they had at least one atom closer than 4.5 Å to the reference molecule. A PDB file containing the reference molecule and all relevant satellite molecules was created. This PDB file was read by the program *COUNT_CONTS* (available from the author upon request) in order to calculate for each of the atoms of the reference molecule the number of non-H atoms within a certain distance threshold. The resulting number is called the atomic contact number ACN. The linear correlation coefficient between the refined ADPs ($B_{\rm refined}$) and the ACN, CC($B_{\rm refined}$, ACN), was also calculated using the program *COUNT_CONTS*. For 30 of the PDB entries, the ACN and the corresponding CC($B_{\rm refined}$, ACN) were calculated based on sphere radii from 1.5 to 10.0 Å in 0.1 Å steps in order to establish the optimal sphere radius for calculating ACNs. Furthermore, a least-squares linear fit of $B_{\rm refined}$ against ACN values was performed, establishing the dependence of the refined ADPs on the ACN. A minimum temperature factor $B_{\rm min}$ was then chosen based on the temperature factors of the atoms exhibiting the highest atomic contact numbers in the structure.

2.3. Prediction of temperature factors

ADPs were predicted using four different models (Table 1) using the program *PREDICT_BS* (available from the author upon request). For the first model, a linear relationship between $B_{\rm refined}$ and ACN was assumed. A least-squares linear fit was performed to determine the y-axis intercept a and the slope b (Table 1). These two numbers were then used to predict ADPs ($B_{\rm predicted\text{-}1}$) according to model 1 (see Table 1). If the predicted ADP turned out to be smaller than $B_{\rm min}$ (see previous paragraph) then it was set to the value of $B_{\rm min}$. For the second model, an inverse relationship between $B_{\rm refined}$ and ACN was assumed and a least-squares linear fit of $B_{\rm refined}$ *versus* $10\,000 \times (\mathrm{ACN}^{-1})$ was performed. $B_{\rm predicted\text{-}2}$ was then calculated according to model 2 (see Table 1). For the third model, an inverse relationship between $B_{\rm refined}$ and the square root of ACN was assumed and a least-squares linear fit of $B_{\rm refined}$ *versus* $1000 \times (\mathrm{ACN}^{-0.5})$ was performed. Again, $B_{\rm predicted\text{-}3}$ was then calculated according to model 3 (see Table 1). For model 4, a Gaussian relationship between $B_{\rm refined}$ and

Table 1
The four ADP models used in this study.

Model	Equation	Condition
1	$B_{\text{predicted-1}} = a + b \times \text{ACN}$	If ($B_{\text{predicted-1}} < B_{\text{min}}$) $B_{\text{predicted-1}} = B_{\text{min}}$
2	$B_{\text{predicted-2}} = a + b \times 10000/\text{ACN}$	If ($B_{\text{predicted-2}} < B_{\text{min}}$) $B_{\text{predicted-2}} = B_{\text{min}}$
3	$B_{\text{predicted-3}} = a + b \times 1000/\text{ACN}^{1/2}$	If ($B_{\text{predicted-3}} < B_{\text{min}}$) $B_{\text{predicted-3}} = B_{\text{min}}$
4	$B_{\text{predicted-4}} = a \times \exp(-b \times \text{ACN}^2) + c$	—

ACN was assumed (model 4, Table 1). The nonlinear least-squares fit of B_{refined} *versus* a function of the type $B = a \times \exp(-b \times \text{ACN}^2) + c$ was carried out using the program *GNUPLOT*. In order to circumvent the problem of overflowing arrays in *GNUPLOT*, the observed values of B_{refined} were first averaged as a function of ACN and the averaged values were then fitted against the ACN values. Although this leads to suboptimal fits, the principle is left unchanged. In contrast to models 1–3, no minimum ADP value was applied in this case, since the equation used produces an inherent minimum value (parameter c) by definition. For each of the four models, the predicted ADPs were then scaled to the refined ADPs by assuming that the average predicted and refined ADPs are identical. In order to mimic the B-factor restraints typically used in refinement, the predicted ADPs were further smoothed by calculating for every atom a smoothed ADP value which comprises the average of the predicted ADP value for this atom and the average of the predicted ADP values of all atoms covalently bound to the atom in question. This leads to a total of eight B-factor models per structure examined. For each structure, ten PDB files were then created with (i) the original refined ADPs per atom (B_{refined}), (ii) the averaged ADP assigned to all atoms, (iii)–(vi) the predicted ADPs per atom ($B_{\text{predicted-1,2,3,4}}$) and (vii)–(x) the predicted and smoothed ADPs per atom ($B_{\text{smoothed-1,2,3,4}}$). As quality measures, Pearson's linear and Spearman's rank order correlation coefficients were calculated between the refined ADPs and the predicted ADPs as well as between the refined and smoothed ADPs. In addition, root-mean-square deviations (RMSDs), mean absolute deviations (MADs) and relative mean absolute deviations (RMADs) as defined by Halle (2002) between B_{refined} on the one side and $B_{\text{predicted}}$ and B_{smoothed} on the other were calculated.

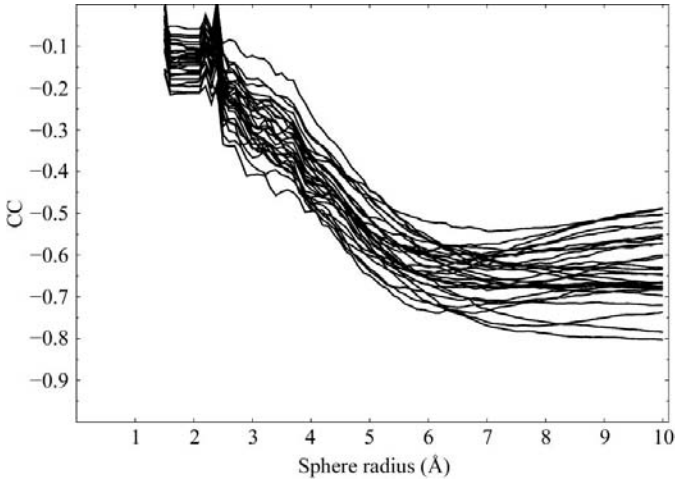

Figure 1
Dependence of the linear correlation coefficient CC(B_{refined}, ACN) on the radius of the sphere used for calculating the ACN for 30 examples from the list of PDB files used in this study.

2.4. Calculation of *R* factors

The CIF files associated with the respective PDB entries were converted to MTZ files using the program *CIF2MTZ* (Collaborative Computational Project, Number 4, 1994). If necessary, the reflections were reduced to the asymmetric unit and re-sorted using *SFTOOLS* (Collaborative Computational Project, Number 4, 1994). The overall Wilson B factor for each of the data sets was calculated using the program *WILSON* (Collaborative Computational Project, Number 4, 1994). For each of the four ADP models per structure, the observed structure-factor amplitudes F_{obs} were then scaled to the calculated amplitudes F_{calc} using all reflections to the maximum resolution possible, a bulk-solvent-type scaling and the six-parameter anisotropic scaling as implemented in the program *REFMAC*5 (Collaborative Computational Project, Number 4, 1994; Murshudov *et al.*, 1997). R factors were then calculated to the maximum resolution possible and to resolutions of 2.0, 2.4 and 3.0 Å also using the program *REFMAC*5.

3. Results and discussion

3.1. The database

The input thresholds chosen resulted in the return of 1310 protein structures with a mean length of 375 amino-acid residues (ranging from 101 to 999 residues) and a mean resolution of 1.64 Å (range 1.50–1.80 Å). The mean R factor of the structures was 17.6% (ranging from 5.0 to 20.0%) and the mean free R factor was 20.9% (ranging from 6.0 to 28.0%). Note that for 76 structures (5.8%) no free R factor had been reported. 219 of the 1310 protein structures (16.7%) had no associated structure-factor file deposited. Consequently, these 219 data sets were excluded, leaving a total of 1091 structures for further examination. A further 126 files failed to be processable using the standard scripts employing the programs *CIF2MTZ*, *SFTOOLS*, *WILSON* and *REFMAC*5. The reasons for this are mostly associated with the CIF files and include the presence of intensities instead of structure-factor amplitudes (52 cases) or extra items present in the CIF files such as, for instance, Hendrickson–Lattman coefficients (21 cases). In a few cases *SFTOOLS* could not handle the output file from *CIF2MTZ* and in some other cases problems arose from nonstandard PDB files (25 cases). Altogether, this left 956 files available for subsequent analysis. These structures

exhibit a mean length of 388 amino-acid residues (ranging from 101 to 999 residues) and a mean resolution of 1.64 Å (range 1.50–1.80 Å). The mean *R* factor of the structures was 17.6% (ranging from 9.0 to 20.0%) and the mean free *R* factor was 21.4% (ranging from 12.0 to 27.0%). This demonstrates that the loss of almost 30% of the originally picked coordinate files did not introduce any sort of bias into the database. With 956 coordinate and structure-factor files remaining, the database is still large enough by far to carry out meaningful statistical calculations and to draw statistically significant conclusions.

3.2. Definition of the size of the sphere for atomic contact number calculation

Fig. 1 shows the dependence of the linear correlation coefficient CC($B_{\rm refined}$, ACN) on the sphere radius used for computing the ACN values for 30 structures taken from the list of PDB entries of this study. Some of the curves exhibit a broad and not well defined minimum at sphere radii from 6 to 8 Å, whereas others appear to asymptotically move towards a minimum value. Based on this figure, the optimal sphere radius to be used for further analysis was defined as 7.0 Å. The value corresponds closely to the value of 7.35 Å used by Halle (2002), which was based on the radial distribution of non-H atoms around C^{α} atoms in protein structures. It is evident from Fig. 1, however, that the choice of 7.0 Å for this parameter is not really critical. Almost any value between 6 and 8 Å can be expected to yield nearly identical results.

3.3. Relationship between refined ADPs and atomic contact number

Fig. 2 shows scatter plots for the three example structures 2jg6, 1jy3 and 2cki of the refined ADPs ($B_{\rm refined}$) and the ACN values calculated based on a sphere of radius 7.0 Å. It is evident from the three parts of the figure that a clear correlation exists between $B_{\rm refined}$ and the ACN. The three corresponding linear correlation coefficients are −0.60, −0.76 and −0.58, respectively. However, it is also evident from Fig. 2 that the relationship between $B_{\rm refined}$ and the ACN is not necessarily linear. For instance, the moduli of the linear correlation coefficients for an inverse relationship between $B_{\rm refined}$ and ACN are practically identical to those listed above.

3.4. Prediction of ADPs

A least-squares linear fit of $B_{\rm refined}$ *versus* ACN yields the dependence between the two, which is depicted by the straight sloping blue line. The horizontal part of the blue line indicates the minimum ADP value allowed for predicting ADPs from the ACN numbers. The corresponding average values for $B_{\rm refined}$ as a function of ACN are shown as the red line. It is clear from the distribution and the shape of the red line that a linear fit may not be the optimal way to describe the dependence. Therefore, three additional *B*-factor models (Table 1)

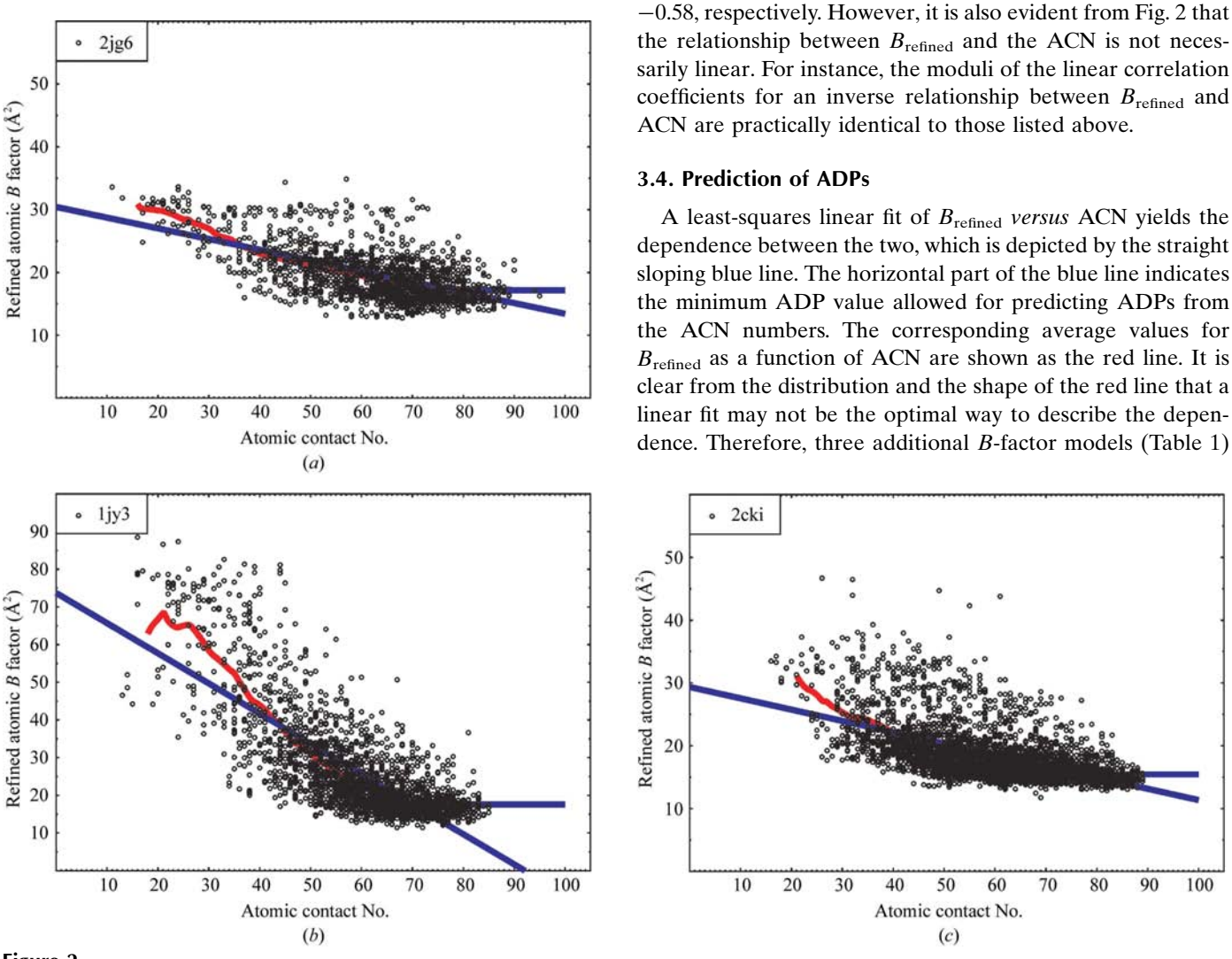

Figure 2
Scatter plots of the refined ADP values $B_{\rm refined}$ *versus* the ACN based on a sphere radius of 7.0 Å for PDB entries 2jg6 (*a*), 1jy3 (*b*) and 2cki (*c*). The red curve shows the averaged ADPs per ACN smoothed over a window of 11; the blue lines show the linear fit of the distribution and the average *B* factors at high ACN numbers. The *y*-axis intercepts, slopes and minimum *B* values for the three cases are 30.43, 73.70 and 29.36 Å^2, −0.17, −0.80 and −0.18 Å^2, and 17.18, 17.54 and 15.45 Å^2, respectively.

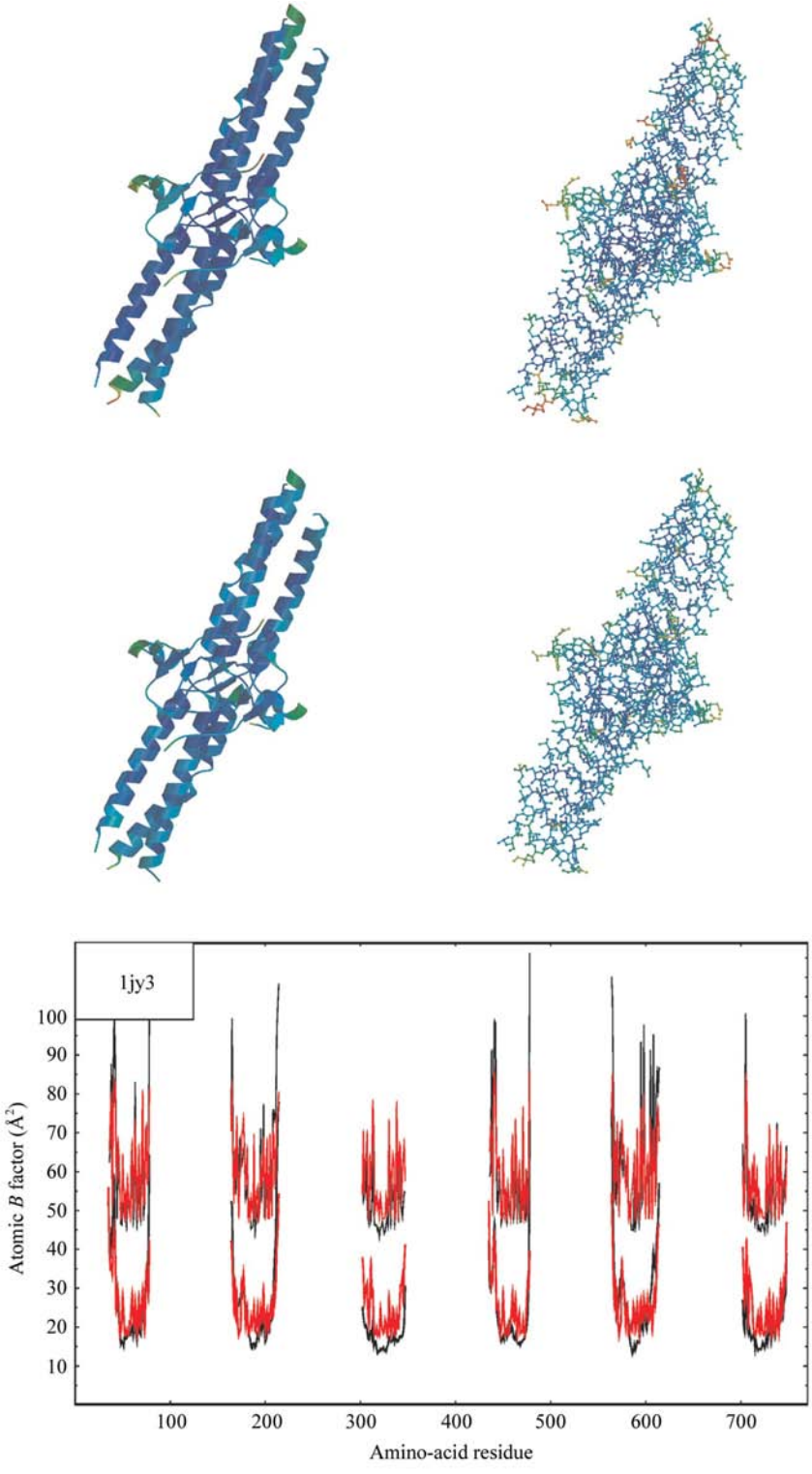

Figure 3
Refined and predicted ADP values for the structure 1jy3 (Madrazo *et al.*, 2001). The top panels show a ribbon plot and an all-atom representation of the structure coloured according to the refined ADPs; the middle panels show the same representations of the structure coloured according to the predicted ADPs. The colouring is as follows: blue for the lowest temperature factors and red for the highest, with rainbow colours in between. The bottom panel shows the refined ADPs (black) and the predicted ADPs (red) for the six protein chains in the asymmetric unit averaged for the main-chain atoms (lower curves) and for the side-chain atoms (upper curves). The side-chain curves have been shifted upwards by 30 Å^2 for clarity.

have been introduced. For each of the four models, the parameters a, b and c (or $B_{\rm min}$) were determined and used to predict atomic temperature factors $B_{\rm predicted\text{-}1,2,3,4}$ according to the equations given in Table 1. Figs. 3 and 4 give a visual comparison of the refined and the predicted (according to model 4, Table 1) ADPs. Fig. 3 displays the structure of a fragment of bovine fibrinogen (Madrazo *et al.*, 2001). It is clearly discernible from the top and the middle panels of Fig. 3 that the essential features of the ADP modulation have been predicted correctly, at least for the main-chain atoms. It is also striking that one of the most mobile side chains (Arg64 of chain *R*; top right panel in Fig. 3, close to the centre of the structure) is not predicted to be flexible although it is very flexible in the structure. The reason for this is unclear. A more quantitative view is shown in the bottom panel. Again, it is evident that the major features have been predicted correctly. However, what can also be noticed is that the extremes of the ADP variation are underpredicted. Whether this is a consequence of the simple model used for prediction or whether this is a sign of over-refinement of the structure cannot be stated with confidence. Fig. 4 shows a view of the structure of ulilysin (Tallant *et al.*, 2006) in the same arrangement. Again, it is evident that the essential features of the ADP modulation are predicted correctly, but that the extremes of flexibility (for instance at the flexible loop between residues 182 and 186 in both subunits at the dimerization interface) are not predicted.

Various global quality indicators for the predicted ADPs are listed in Table 2. Most of the indicators favour the ADPs predicted by model 4, although the difference from models 1, 2 and 3 is rather small and probably not really significant. Spearman's rank order correlation coefficient, which was used by Halle (2002), appears to be unable to distinguish between the different models. The average correlation coefficient between refined and predicted temperature factors is 0.67, which indicates that the general features of the *B*-factor modulation are correctly

predicted in most cases. The highest observed correlation coefficient is 0.85, pointing to an essentially correct prediction. The average correlation coefficient of 0.67 is slightly higher than that reported by Kundu *et al.* (2002), who obtained 0.65 for the GNM model approach when the neighbouring molecules in the crystal were included. This shows that the model based on local packing density is somewhat superior to the GNM model, although it is computationally much less demanding. It also shows that the neighbouring molecules in the crystal lattice have a significant influence on the quality of the prediction. Halle (2002) also obtained an average correlation coefficient of 0.67 for 38 high-quality protein structures, but again only when the effect of the crystal neighbours was taken into account. A further interesting quality indicator to look at is the RMAD as defined by Halle (2002). A comparison of two identical sets of numbers would produce an RMAD value of 0.0, whereas fitting any set of numbers by the average value yields an RMAD of 1.00. The fact that for each of the four models an average RMAD of about 0.80 is obtained here means that about 20% of the ADP variation observed in structures in which individual ADPs have been refined can be correctly predicted. In extreme cases this can go up to 50%, as indicated by RMAD values of about 0.50.

3.5. Smoothing of *B* factors

The smoothing applied to the predicted ADPs of models 1–4 yielded a slight improvement in the quality indicators (Table 3). Although this improvement is rather small, it is consistent throughout all of the structures examined and shows up in all quality indicators used. Since macromolecular structures are typically refined using *B*-factor restraints (across one or two bonds), the observed improvement may just be a consequence of mimicking the utilization of such restraints in refinement.

3.6. Usefulness in structure refinement

Table 4 displays the crystallographic *R* factors which can be obtained for the various ADP models. As a reference for what can be achieved, the *R* factors against a structure with the original refined ADP values are also given. Since all solvent molecules or cofactors, substrates, inhibitors *etc.* have been stripped from the protein molecules, the *R* factors are different from those in the respective PDB entries. The upper *R*-factor boundaries are defined by the *R* factors which are obtained when only the average ADP value is used for all atoms. From the numbers presented in Table 4, it is obvious that with the sets of ADPs derived from models 1–4, *R* factors can be obtained which are about halfway between those based on average ADPs and those based on individually refined ADPs. It is also noteworthy that the four models employed do not differ much,

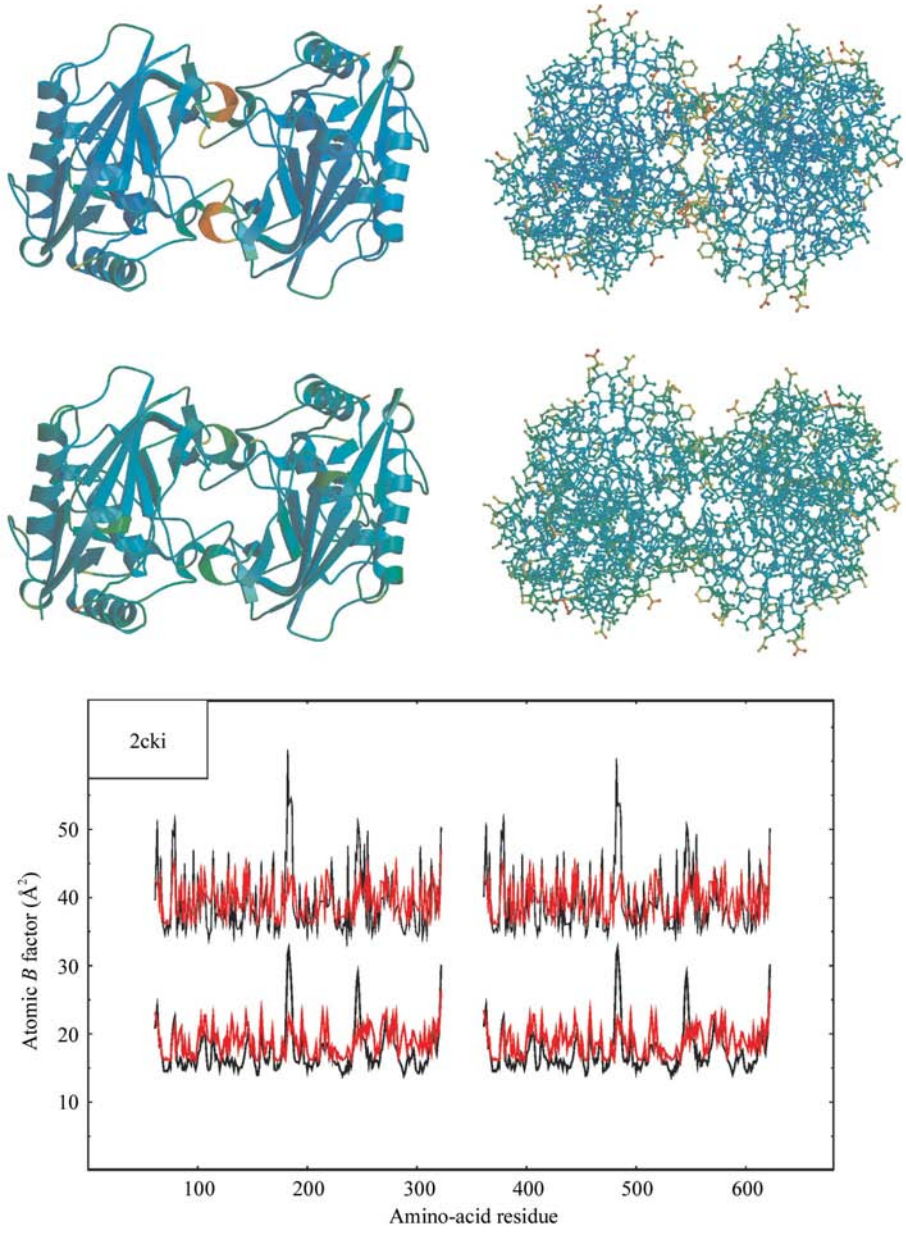

Figure 4
Refined and predicted ADP values for the structure 2cki (Tallant *et al.*, 2006). The top panels show a ribbon plot and an all-atom representation of the structure coloured according to the refined ADPs; the middle panels show the same representations of the structure coloured according to the predicted ADPs. The colouring is as in Fig. 3. The bottom panel shows the refined ADPs (black) and the predicted ADPs (red) for the two protein chains in the asymmetric unit averaged for the main-chain atoms (lower curves) and for the side-chain atoms (upper curves). The side chain curves have been shifted upwards by 20 Å^2 for clarity.

Table 2
Quality indicators for the predicted ADP values based on models 1–4.

	Mean value	Standard deviation	Maximum value	Minimum value
Linear correlation coefficient CC($B_{\rm refined}$, $B_{\rm predicted}$)				
Model 1	0.659	0.073	0.819	0.278
Model 2	0.631	0.084	0.814	0.136
Model 3	0.656	0.080	0.827	0.186
Model 4	0.670	0.077	0.848	0.276
Rank order correlation coefficient				
Model 1	0.627	0.073	0.813	0.299
Model 2	0.626	0.074	0.813	0.298
Model 3	0.626	0.073	0.813	0.298
Model 4	0.625	0.074	0.813	0.298
Root-mean-square deviation RMSD (Å^2)				
Model 1	5.85	2.20	17.05	2.18
Model 2	6.06	2.29	17.37	2.27
Model 3	5.87	2.24	17.21	2.23
Model 4	5.78	2.19	17.11	2.24
Mean absolute deviation MAD (Å^2)				
Model 1	4.42	1.55	13.66	1.82
Model 2	4.51	1.61	13.63	1.85
Model 3	4.41	1.57	13.55	1.84
Model 4	4.35	1.54	13.40	1.88
Relative mean absolute deviation RMAD				
Model 1	0.79	0.08	1.53	0.57
Model 2	0.81	0.07	1.50	0.61
Model 3	0.79	0.08	1.50	0.57
Model 4	0.78	0.08	1.45	0.51

Table 3
Quality indicators for the predicted and smoothed ADP values based on models 1–4.

	Mean value	Standard deviation	Maximum value	Minimum value
Linear correlation coefficient CC($B_{\rm refined}$, $B_{\rm smoothed}$)				
Model 1	0.671	0.075	0.834	0.287
Model 2	0.650	0.084	0.833	0.133
Model 3	0.671	0.081	0.843	0.184
Model 4	0.683	0.078	0.864	0.285
Rank order correlation coefficient				
Model 1	0.638	0.076	0.824	0.309
Model 2	0.638	0.076	0.825	0.308
Model 3	0.638	0.076	0.825	0.308
Model 4	0.638	0.076	0.825	0.309
Root-mean-square deviation RMSD (Å^2)				
Model 1	5.76	2.19	17.00	1.69
Model 2	5.76	2.19	17.00	1.69
Model 3	5.74	2.22	17.19	1.63
Model 4	5.66	2.19	17.14	1.59
Mean absolute deviation MAD (Å^2)				
Model 1	4.35	1.54	13.62	1.40
Model 2	4.44	1.61	13.70	1.35
Model 3	4.33	1.57	13.45	1.34
Model 4	4.27	1.55	13.51	1.29
Relative mean absolute deviation RMAD				
Model 1	0.78	0.07	1.07	0.56
Model 2	0.79	0.07	1.08	0.60
Model 3	0.77	0.07	1.05	0.56
Model 4	0.76	0.08	1.07	0.50

although model 4 appears to be slightly better than the other three. This corroborates the findings based on the correlation coefficients CC($B_{\rm refined}$, $B_{\rm predicted}$) displayed in Table 2 and the visual impression from Fig. 2, where it seems that an exponential fit would be better for describing the distribution at lower ACN values than a linear fit. Table 5 shows the relative improvement of the four three-parameter models over the one-parameter model, which just takes into account the average ADP value. It is clear from the numbers that an ADP modulation which accounts for about 50% of the improvement in R factor over the use of an average ADP value for each atom can be predicted solely based on the local packing density of the atom. This fraction increases slightly at lower resolution. It is important to note that the parameters a, b and c (or $B_{\rm min}$) are variable and that they are a function of the structure example studied, although significant correlations between the experimentally derived Wilson B factors and the parameters a and c of the four models exist. It may therefore be possible to derive the parameters a and c from the Wilson plot and to use the parameter b as the only refinable parameter. In summary, it may well be the case that refining just one or possibly three parameters instead of individual ADPs may turn out to be a sensible approach for macromolecular structure refinement, especially at lower resolution.

Table 4
Crystallographic R factors (%) in various resolution ranges (Å).

Given are the mean values and the standard deviations over the 956 structures used in this study.

	20.0–$d_{\rm min}$	20.0–2.0	20.0–2.4	20.0–3.0
Original ADP values	26.23 (2.81)	25.70 (2.73)	25.93 (2.84)	23.84 (2.83)
Average ADP values	29.98 (2.56)	28.91 (2.65)	28.41 (2.77)	25.79 (2.81)
Model 1 – predicted	28.21 (2.61)	27.32 (2.64)	27.10 (2.78)	24.80 (2.79)
Model 1 – smoothed	28.18 (2.62)	27.29 (2.64)	27.09 (2.78)	24.78 (2.78)
Model 2 – predicted	28.21 (2.66)	27.36 (2.68)	27.15 (2.82)	24.83 (2.81)
Model 2 – smoothed	28.22 (2.62)	27.36 (2.67)	27.11 (2.81)	24.78 (2.80)
Model 3 – predicted	28.13 (2.65)	27.26 (2.67)	27.07 (2.81)	24.77 (2.81)
Model 3 – smoothed	28.12 (2.64)	27.26 (2.67)	27.04 (2.81)	24.76 (2.81)
Model 4 – predicted	28.11 (2.64)	27.22 (2.67)	27.02 (2.79)	24.75 (2.81)
Model 4 – smoothed	28.08 (2.65)	27.21 (2.66)	26.98 (2.79)	24.74 (2.79)

3.7. Limitations of the approach

There are three principal reasons which will result in incorrect predictions of ADPs by the proposed models. The first is if the ADPs are systematically altered by a TLS motion of the whole molecule or domains of the whole molecule. In such a case, the refined ADPs will contain the effects of the TLS motion(s) and the contribution inherent from the structure. The obvious solution to this is of course to identify the relevant rigid groups in the molecule, for instance following the *TLSMD* approach of Painter & Merritt (2006), perform TLS refinement against the observed diffraction data and eliminate the TLS contribution to the ADPs computationally. The second reason is that if some coordinates are modelled incorrectly the resulting ACNs will be incorrect and consequently the predicted ADPs will be erroneous. Consequently, during the course of structure refinement, when the coordinates are updated the ACNs have to be updated as well in order to ensure the best possible prediction. The third reason is missing coordinates. In the approach presented here, all cofactor and ligand molecules as well as water molecules bound to the surface of the protein have been neglected, although they should in principle contribute to the ACN. An

Table 5
Absolute and relative (in parentheses) improvement of the crystallographic R factors (%) for different resolution ranges (Å) when going from the one-parameter model (average ADPs) to the three-parameter models discussed in this study.

The crystallographic R factors calculated against the coordinates with the original ADPs are taken as the reference R factor.

	20.0–$d_{\rm min}$	20.0–2.0	20.0–2.4	20.0–3.0
Model 1 – predicted	1.77 (47.2)	1.59 (49.5)	1.31 (52.8)	0.99 (50.8)
Model 1 – smoothed	1.80 (48.0)	1.62 (50.5)	1.32 (53.2)	1.01 (51.8)
Model 2 – predicted	1.77 (47.2)	1.55 (48.3)	1.26 (50.8)	0.96 (49.2)
Model 2 – smoothed	1.76 (46.9)	1.55 (48.3)	1.30 (52.4)	1.01 (51.8)
Model 3 – predicted	1.85 (49.3)	1.65 (51.4)	1.34 (54.0)	1.02 (52.3)
Model 3 – smoothed	1.86 (49.6)	1.65 (51.4)	1.37 (55.2)	1.03 (52.8)
Model 4 – predicted	1.87 (49.9)	1.69 (52.6)	1.39 (56.0)	1.04 (53.3)
Model 4 – smoothed	1.90 (50.7)	1.70 (53.0)	1.43 (57.7)	1.05 (53.8)

extended approach taking all those atoms into account should therefore yield even better predictions.

3.8. Usefulness in validation

As discussed in §3.7, one reason for a local deviation of the refined ADP of an atom or a group of atoms from their predicted ADPs is that the atoms may be incorrectly modelled. By investigating the local discrepancies between $B_{\rm refined}$ and $B_{\rm predicted}$, one may therefore be able to pinpoint problematic regions in the structure during model building and refinement. Alternatively, this approach may be used for structure-validation purposes.

4. Conclusions and outlook

Based on the data presented, it is clear that atomic coordinates and atomic displacement parameters should not be considered to be independent variables in protein-structure refinement and analysis. Based on the architecture of a protein structure, the atomic flexibilities expressed in terms of ADP values can be predicted qualitatively. Three additional parameters per structure are then required for a semi-quantitative prediction, which is able to explain 50% of the improvement in crystallographic R factor on the way from a single-parameter average ADP value structure to a structure with individually refined ADPs. Especially at lower resolution, this may constitute an attractive alternative for individual ADP refinement.

The models described for ADP prediction are extremely simple and only rely on the local packing density. It may well be possible to improve the prediction accuracy by taking the actual structure in terms of covalent and noncovalent interactions into account. A further natural extension of the proposed approach is to consider all resolved atoms in a crystal structure including cofactor, ligand and water atoms. Also, the proposed models for predicting ADPs may be combined, for instance with TLS refinement. Since the contribution of TLS to the total value of an ADP originates from a different source than that described here, a combined approach may be the key to even better predictions.

Last, but not least, yet another possibility for extension of the method is to predict anisotropic ADP values. So far, the approach has been limited to isotropic ADPs. However, since the local packing density is anisotropic, it may well be possible to predict anisotropic temperature factors as well. This calls for further exploration in the future.

I would like to thank Venkataraman Parthasarathy and Dr Santosh Panjikar (EMBL Hamburg) for their help in preparing the script for carrying out this work and Dr Oliviero Carugo (University of Pavia) for providing his program *CPC*. Furthermore, I would like to thank Dr Victor Lamzin (EMBL Hamburg) for his linear fitting routine and the members of Victor's group for a helpful discussion on this topic. Finally, I would like to thank the two anonymous referees for some very valuable suggestions.

References

Bahar, I., Atilgan, A. R., Demirel, M. C. & Erman, B. (1998). *Phys. Rev. Lett.* **80**, 2733–2736.
Bahar, I., Atilgan, A. R. & Erman, B. (1997). *Fold. Des.* **2**, 173–181.
ben-Avraham, D. & Tirion, M. M. (1998). *Physica A*, **249**, 415–423.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Carugo, O. & Argos, P. (1999). *Acta Cryst.* D**55**, 473–478.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Gohlke, H., Kuhn, L. A. & Case, D. A. (2004). *Proteins*, **56**, 322–337.
Haliloglu, T. & Bahar, I. (1999). *Proteins*, **37**, 654–667.
Haliloglu, T., Bahar, I. & Erman, B. (1997). *Phys. Rev. Lett.* **79**, 3090–3093.
Halle, B. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 1274–1279.
Higo, J. & Umeyama, H. (1997). *Protein Eng.* **10**, 373–380.
Hinsen, K. & Kneller, G. R. (1999). *J. Chem. Phys.* **111**, 10766–10769.
Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. (2001). *Proteins*, **44**, 150–165.
Kundu, S., Melton, J. S., Sorensen, D. C. & Phillips, G. N. Jr (2002). *Biophys. J.* **83**, 723–732.
Levitt, M., Sander, C. & Stern, P. S. (1985). *J. Mol. Biol.* **181**, 423–447.
MacKerell, A. D. Jr *et al.* (1998). *J. Phys. Chem.* **102**, 3586–3616.
Madrazo, J., Brown, J. H., Litvinovich, S., Dominguez, R., Yakovlev, S., Medved, L. & Cohen, C. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 11967–11972.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Painter, J. & Merritt, E. A. (2006). *Acta Cryst.* D**62**, 439–450.
Schomaker, V. & Trueblood, K. N. (1968). *Acta Cryst.* B**24**, 63–76.
Tallant, C., Garcia-Castellanos, R., Seco, J., Baumann, U. & Gomis-Ruth, F. X. (2006). *J. Biol. Chem.* **281**, 17920–17928.
Tirion, M. M. (1996). *Phys. Rev. Lett.* **77**, 1905–1908.
Wang, G. & Dunbrack, R. L. Jr (2003). *Bioinformatics*, **19**, 1589–1591.