Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the precision of calculated solvent-accessible surface areas

Marian Novotny,[a,b] Marvin Seibert[a] and Gerard J. Kleywegt[a]*

[a]Department of Cell and Molecular Biology, Uppsala University, Biomedical Centre, Box 596, SE-751 24, Uppsala, Sweden, and [b]Linnaeus Centre for Bioinformatics, Biomedical Centre, Box 598, SE-751 24, Uppsala, Sweden

Correspondence e-mail: gerard@xray.bmc.uu.se

Received 29 August 2006
Accepted 23 October 2006

The fact that protein structures are dynamic by nature and that structure models determined by X-ray crystallography, electron microscopy (EM) and nuclear magnetic resonance (NMR) spectroscopy have limited accuracy limits the precision with which derived properties can be reported. Here, the issue of the precision of calculated solvent-accessible surface areas (ASAs) is addressed. A number of protein structures of different sizes were selected and the effect of random shifts applied to the atomic coordinates on ASA values was investigated. Standard deviations of the ASA calculations were found to range from ∼10 to ∼80 Å^2. Similar values are obtained for a handful of cases in the Protein Data Bank (PDB) where 'ensembles' of crystal structures were refined against the same data set. The ASA values for 69 hen egg-white lysozyme structures were calculated and a standard deviation of the ASA of 81 Å^2 was obtained (the average ASA value was 6571 Å^2). The calculated ASA values do not show any correlation with factors such as resolution or overall temperature factors. A molecular-dynamics (MD) trajectory of lysozyme was also analysed. The ASA values during the simulation covered a range of more than 800 Å^2. If different programs are used, the ASA values obtained for one small protein show a spread of almost 600 Å^2. These results suggest that in most cases reporting ASA values with a precision better than 10 Å^2 is probably not realistic and a precision of 50–100 Å^2 would seem prudent. The precision of buried surface-area calculations for complexes is also discussed.

1. Introduction

The solvent-accessible surface of a protein is the area where interactions between the protein and solvent and solutes take place. This surface is usually calculated by rolling a sphere (representing a solvent molecule) over the van der Waals surface of the protein. The size of the area covered by the centre of this sphere is called the solvent-accessible surface area (ASA) of the protein (Lee & Richards, 1971). The ASA concept was first introduced to study how hydrophobic amino acids behave during protein folding (Lee & Richards, 1971). Since then, ASA measurements have found widespread application, *e.g.* to predict the oligomerization state of proteins (Henrick & Thornton, 1998), to determine the preferred ground state of an enzyme and assess the role of active-site residues (Mazumder-Shivakumar & Bruice, 2005), to improve the ranking of docking solutions (Duan *et al.*, 2005), to predict the structure of surface loops in proteins (Das & Meirovitch, 2003), to define the side-chain conformational entropy at the interaction surfaces of proteins (Cole & Warwicker, 2002), to characterize protein–nucleic acid recognition sites (Nadassy *et al.*, 1999) and to calculate absolute free energies of binding of ligands to a protein in molecular-dynamics simulations (Bartels *et al.*, 2005).

Calculated ASA values are usually reported without an estimate of the error bar and with a precision that varies between 0.1 and 100 Å^2. There are a number of factors that limit the precision with which ASA values can realistically be reported. Firstly, the accuracy with which X-ray, electron microscopy (EM) and nuclear magnetic resonance (NMR) structures are determined is limited. The coordinate error of typical crystal structures is likely to vary from ∼0.05 Å at

Table 1
Effect of random perturbations on ASA values calculated for proteins of different sizes.

$\langle\sigma_{ASA}\rangle$ is the average standard deviation; the standard deviation of the ASA values calculated for each of ten randomly perturbed models was determined and their values were averaged over five proteins of similar size. $\langle\Delta_{ASA}\rangle$ is the average ASA difference between the original and perturbed models for all 5 × 10 models generated for each protein size and each perturbation amplitude.

Shift (Å)	0.01		0.05		0.1		0.2		0.5	
Residues	$\langle\sigma_{ASA}\rangle$ (Å^2)	$\langle\Delta_{ASA}\rangle$ (Å^2)	$\langle\sigma_{ASA}\rangle$ (Å^2)	$\langle\Delta_{ASA}\rangle$ (Å^2)	$\langle\sigma_{ASA}\rangle$ (Å^2)	$\langle\Delta_{ASA}\rangle$ (Å^2)	$\langle\sigma_{ASA}\rangle$ (Å^2)	$\langle\Delta_{ASA}\rangle$ (Å^2)	$\langle\sigma_{ASA}\rangle$ (Å^2)	$\langle\Delta_{ASA}\rangle$ (Å^2)
∼100	7	4	13	11	19	12	24	13	41	62
∼200	9	4	20	6	30	23	39	23	64	76
∼300	9	7	21	17	30	17	39	37	72	71
∼500	13	4	25	10	28	22	45	52	81	210
Average	10	5	20	11	27	19	37	31	65	105

very high resolution to >0.5 Å at very low resolution (Cruickshank, 1999). The accuracy can be even lower if the model contains non-random errors (Kleywegt, 2000). Secondly, protein structures are dynamic and therefore any derived properties (such as ASA values) are bound to fluctuate with time. A crystal structure is but one snapshot and some details of it may even be irrelevant under physiological conditions (*e.g.* owing to crystal-packing interactions or non-physiological pH or ionic strength).

In this contribution, the effect of random errors and possible non-random phenomena as well as of protein dynamics on the precision of calculated ASA values is assessed. (Since 'true' ASA values are unknown, it is impossible to assess their accuracy.) Given the difficulty of obtaining coordinate-error estimates for individual atoms and the propagation of these errors into the calculated ASA values (since ASA values cannot be calculated analytically from the coordinates), randomly perturbed X-ray structures were used to study the effect of random errors, multiple independent structure determinations were used to study the combined effect of random errors and possible non-random phenomena and a molecular-dynamics simulation was used to study the effect of dynamic fluctuations.

2. Methods

Three data sets were used in this study. Data set 1 was a set of 20 structures of moderate resolution (1.8–2.2 Å) that was created by randomly picking five structures from the Protein Data Bank (PDB; Berman *et al.*, 2000) with a length of ∼100 amino acids (PDB codes 1bxu, 1mgw, 1yd4, 1ybz, 1vku), five of ∼200 amino acids (1hyl, 1tev, 1y88, 1ywm, 1z77), five of ∼300 amino acids (1aq1, 1b38, 1e5j, 1h0v, 1ofc) and five of ∼500 amino acids (1jxj, 1mfv, 1nm9, 1g4n, 1smd). *MOLEMAN*2 (Kleywegt, 1997) was used to generate perturbed models of each structure in the first data set by applying random shifts to all Cartesian coordinates. These shifts were taken from a uniform distribution with an amplitude of 0.01, 0.05, 0.1, 0.2 and 0.5 Å, resulting in average positional shifts of approximately the same magnitude as the corresponding amplitude. Ten different randomly perturbed models were generated at each of these five amplitudes for every structure in this data set. For each set of ten models, the standard deviation of the ASA values was calculated, as was the average difference between the ASA values and that calculated for the unperturbed model. The results were then averaged for each set of five proteins of similar size and for each perturbation amplitude (Table 1).

Data set 2 consisted of 69 hen egg-white lysozyme (HEWL) structures that all contain an identical number of atoms (1001, after deleting any and all water molecules and hetero compounds and ignoring all but the first occurrence of any atoms with alternative conformations) for which structure factors had been deposited and that were all determined at 1.1–2.3 Å resolution. Of these structures, 52 had been crystallized in space group $P4_32_12$, ten in $P2_12_12_1$, three in $P2_1$, three in $P1$ and one in $C2$. The ASA values of these proteins were calculated and correlations with various properties were investigated, namely with the resolution of the study, with the Wilson B factor and the average protein B factor (these statistics were extracted from the Electron Density Server; EDS; Kleywegt *et al.*, 2004) and with the all-atom root-mean-square distance (r.m.s.d.) of each model to the reference structure (PDB code 1aki; Artymiuk *et al.*, 1982), which was calculated with *LSQMAN* (Kleywegt, 1996).

Data set 3 consisted of 2000 structures taken from a molecular-dynamics (MD) trajectory of hen egg-white lysozyme. The simulation was performed with the *GROMACS* 3.3 MD package (van der Spoel *et al.*, 2005) as described by Patriksson *et al.* (2007). Briefly, the protein structure (PDB code 1aki; Artymiuk *et al.*, 1982) was simulated with the OPLS-AA force field (Kaminski *et al.*, 2001) in explicit water (8499 TIP4P molecules) and with eight counter-ions to neutralize the charge. Berendsen weak coupling (Berendsen *et al.*, 1984) was used to maintain the temperature at 300 K and the pressure at 1 bar (10^5 Pa). The computation of short-range interactions was truncated with a 9/14 Å twin-range cutoff, whereas long-range electrostatics were calculated with the smooth particle-mesh Ewald algorithm (Darden *et al.*, 1993; Essmann *et al.*, 1995). H atoms were treated as virtual particles. The simulation used time steps of 4 fs and structures were saved every 10 ps. The total simulation time was 20 ns.

All solvent-accessible surface areas in this study were calculated with the *AREAIMOL* program from the *CCP*4 package (Collaborative Computational Project, Number 4, 1994) using default settings. In the case of alternative conformations, *AREAIMOL* only uses the first of these encountered in the input PDB file.

3. Results

3.1. Effect of random errors

Even structures at atomic resolution contain uncertainties in their atomic coordinates. To investigate their effect on calculated ASA values, 50 randomly perturbed models were generated for each of 20 protein crystal structures. These structures were of specific sequence lengths (five with a length of ∼100 residues, five of ∼200, five of ∼300 and five of ∼500 residues) to enable assessment of the size-dependence of the results. The effect of random errors was modelled by applying random shifts of varying amplitude to all atomic coordinates and the results are summarized in Table 1. The average standard deviations of the calculated ASA values and the average ASA discrepancies (between perturbed and unperturbed models) were found to be of the same order of magnitude (∼10–100 Å^2) and both increase with increasing sequence length and increasing amplitude of

the random shifts, although the effect of the latter was more pronounced.

3.2. Effect of structural heterogeneity

To investigate the effect of possible non-random factors (*e.g.* lattice-dependent crystal-packing contacts, possible modelling errors, different experimental conditions, different refinement protocols *etc.*), ASA values were calculated for 69 HEWL structures with an identical number of atoms. The average ASA value was 6571 Å^2 with a standard deviation of 81 Å^2. The minimum ASA value was 6392 Å^2 and the maximum value was 6836 Å^2 (*i.e.* a range of 444 Å). Fig. 1 shows the relationship between the ASA values and the all-atom r.m.s.d. to the reference structure (1aki), the resolution of the structure, the Wilson *B* factor and the average temperature factor of the model, respectively. The linear correlation coefficients are all close to zero, suggesting that there is no correlation between the ASA values and any of these statistics. In addition, crystal packing does not appear to have a systematic effect on the calculated ASA values.

3.3. Effect of dynamic fluctuations

Under physiological conditions, proteins are in constant motion. However, structure-determination methods capture only one or a few snapshots of all the conformations that a protein can assume. The results of an MD simulation of hen egg-white lysozyme were used to assess how the ASA value of this protein fluctuates during short time periods. Fig. 2 shows the ASA values for 2000 structures along the 20 ns simulation. The average ASA value was 7036 Å^2, with a standard deviation of 114 Å^2. The minimum observed ASA was 6577 Å^2 and the maximum observed value was 7391 Å^2 (*i.e.* a range of 814 Å^2).

4. Discussion

Solvent-accessible surface-area calculation results are widely used to describe and analyse protein structures and to design further experiments. Despite their importance, the precision of these calculations has received little attention to date. Kaliannan *et al.* (1998) have assessed the precision of ASA calculations for individual atom types. They found that the average variation of the ASA for typical atom types lies in the range 2–28 Å^2, with C and O atoms showing larger variations than N and S atoms.

In this work, a number of data sets have been used to obtain some indication of the limits on the precision with which ASA values can reasonably be reported. 50 randomly perturbed models were generated for each of 20 structures by applying random shifts to all atoms. As shown in Table 1, the average standard deviation of the ASA values for these structures ranged from 7 Å^2 for small proteins (around 100 residues) and small shifts (0.01 Å) to 81 Å^2 for large proteins (500 residues) and large shifts (0.5 Å). The standard deviations of the ASA values and the average ASA differences between the crystal structures and the perturbed models were of similar

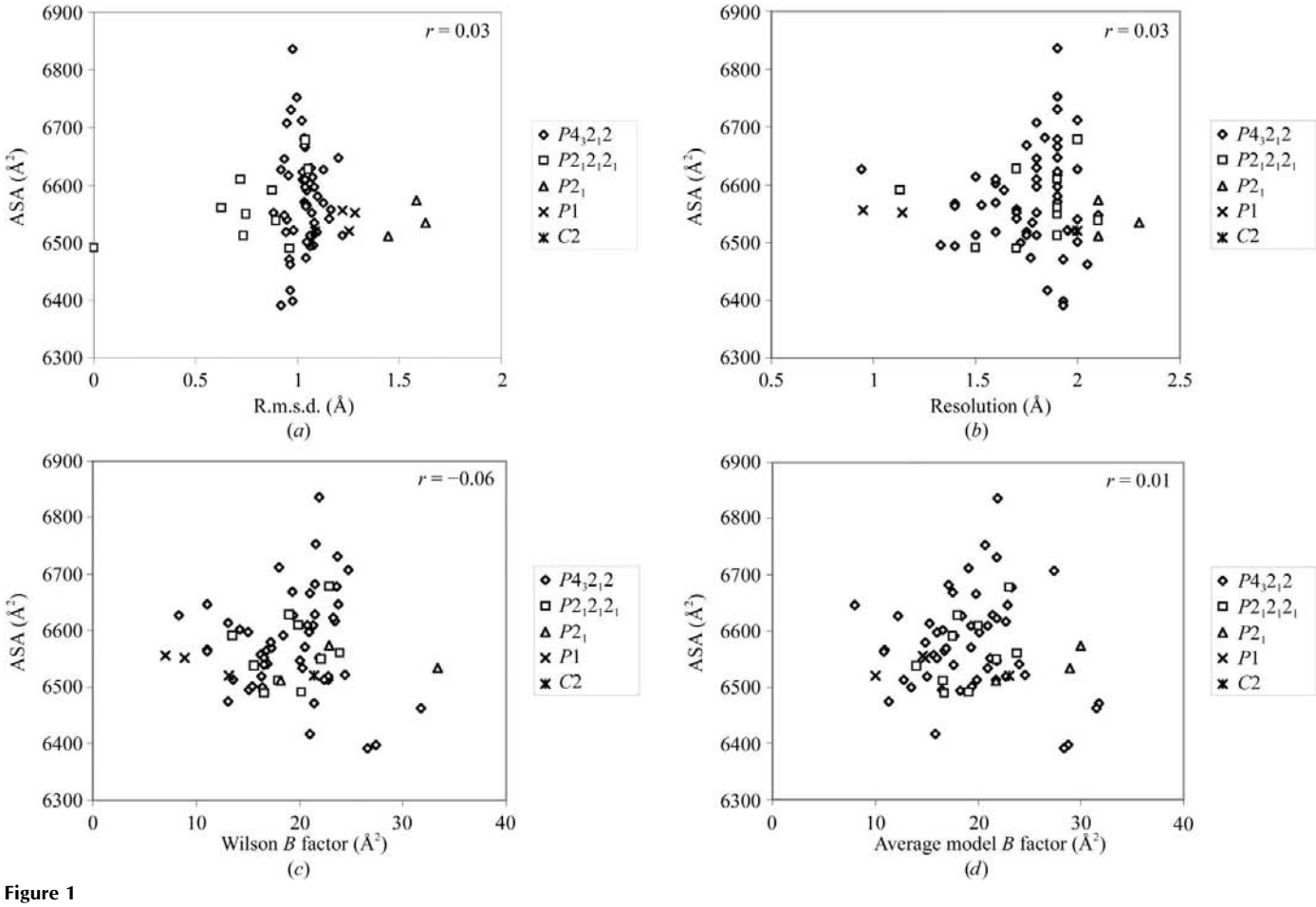

Figure 1
Variation of ASA values calculated for 69 lysozyme structures with (*a*) all-atom r.m.s.d. to the reference structure (1aki), (*b*) resolution, (*c*) Wilson temperature factor and (*d*) average model temperature factor. The value of the linear correlation coefficient (*r*) is shown in each plot. The plots show that the effects of random errors and non-random structural heterogeneity on ASA values are not correlated with any of these parameters.

Table 2
Results of ASA calculations for crystal structures in the PDB that contain more than four models refined against the same data set.

PDB code	No. of models	No. of residues	Resolution (Å)	Average ASA (Å^2)	Standard deviation of ASA (Å^2)	Reference
2d6b	10	129	1.25	6472	13	Ondracek & Mesters (2006)
1hc0	10	129	1.82	6380	16	Ondracek *et al.* (2005)
2ull	16	198	1.5	7833	52	Rader & Agard (1997)
1htq	10	477	2.4	22572	162	Gill *et al.* (2002)

magnitude, suggesting that the precision (standard deviation) is a reasonable measure of the 'accuracy' of the calculations. This is not true for large shifts applied to large proteins, where the average ASA difference is three times greater than the average standard deviation. However, for large shifts the random-perturbation method is probably inadequate as a method for modelling random fluctuations, since the covalent geometry will be severely distorted and van der Waals clashes will be the rule rather than the exception. To check whether the random-perturbation method yields reasonable standard deviations, all crystal structures in the PDB for which multiple models refined against the same data set have been deposited were identified. There are over 60 such entries (October 2006), but most of these contain only two models. Four entries contain more than four models and the standard deviations of their ASA values were calculated (Table 2). Although there are too few observations to draw any firm conclusions, the standard deviations follow the same pattern and are of the same order of magnitude as those calculated with the random-perturbation method (Table 1). It should be noted that the rather large discrepancy between the average ASA values of the two related lysozyme structures 2d6b and 1hc0 mostly arises from the fact that the former model contains a handful of atoms that have not been modelled in the latter. In general, if even a single bulky surface side chain has not been modelled, the effect on the total calculated ASA value can easily be of the order of 100 Å^2, *i.e.* considerably larger than the standard deviations calculated for small proteins.

For the ASA calculations on the set of 69 experimentally solved lysozyme structures (representing five different space groups), a standard deviation of 81 Å^2 was obtained. The difference between the lowest and the highest ASA value was 444 Å (more than five times the standard deviation). This may be explained in part by differences in crystallization and other experimental conditions, but some variation is probably also introduced during the model-building process. No correlation was observed between the ASA values and statistics such as resolution and average *B*-factor values (Fig. 1), which suggests that reasonable estimates of ASA values can be obtained even at low resolution. In this particular case, crystal-packing effects did not appear to have any systematic effect on the calculated ASA values (Fig. 1).

The ASA values fluctuated substantially during a 20 ns MD simulation. The difference between the highest and the lowest recorded ASA value was 814 Å^2 (Fig. 2). The timescale of this simulation is short in comparison with the catalytic reaction timescale or even the timescale required for binding of a substrate. Systematic conformational changes arising from the mode of action of the enzyme are therefore not captured in this simulation. The simulation most likely presents a reasonable picture of the enzyme in its resting state and the magnitude of the ASA fluctuations therefore represents a lower limit to the fluctuations in a native protein of this size. The ASA values of the structures in the simulation were higher than that of the crystal structure. This is probably a consequence of the higher water content in a simulation environment compared with a crystal environment. The simulation favours protein–solvent hydrogen bonds, which may 'open' the structure in the simulation to accommodate more protein–solvent interactions. The transition from the native crystal state to the simulation environment was reflected by increasing ASA values in the first 5 ns of the simulation (Fig. 2).

In this work, consistent use has been made of one program for ASA calculations (*AREAIMOL*) to eliminate the effects of using different parameters, algorithms and programs. However, there are many different programs and servers available to calculate ASA values. Seven of these were tested using a plastocyanin structure (PDB code 1bxu; 99 amino acids; Inoue *et al.*, 1999) to obtain an impression of the effects of using different programs. Default settings were used for all of them and the calculated ASA values are reported with the precision given by the programs. The Vadar server (Willard *et al.*, 2003) calculated a value of 4662.2 Å^2, the *NACCESS* program (Hubbard & Thornton, 1993) 4664.4 Å^2, the *AREAIMOL* program 4673.0 Å^2, the *DSSP* program (Kabsch & Sander, 1983) 4758 Å^2, the POPS server (Cavallo *et al.*, 2003) 4826.0 Å^2, the RPBS ASA server (Richmond, 1984) 4950.0 Å^2 and the StrucTools server running the *MSMS* program (Sanner *et al.*, 1996) 5224.8 Å^2. These results show that ASA values that are calculated by different programs cannot be compared directly. Therefore, when ASA values are reported it is important that the program that was used to calculate them is mentioned.

ASA values are often reported for the interface area buried upon the formation of a complex. For example, if two proteins *A* and *B* form a complex *AB*, the buried surface area (BSA) can be calculated as: BSA = ASA(*A*) + ASA(*B*) − ASA(*AB*). If the coordinates of the individual structures *A* and *B* are taken from the complex *AB*, the errors for the non-interface atoms will be identical and thus cancel out. Therefore, the error in the buried surface area will effectively be of the same order of magnitude as that of a much smaller protein. Moreover, the errors in the ASA values are likely to be distributed non-uniformly over the surface of a protein complex, with residues in the interface being more rigid and therefore having a smaller error bar on their ASA. For these reasons, buried surface areas can be reported with a higher precision than the total ASA of the complex and the individual proteins. However, this is not true if the buried surface area is calculated using different structures (of the two proteins in isolation and of the complex). In this case, error propa-

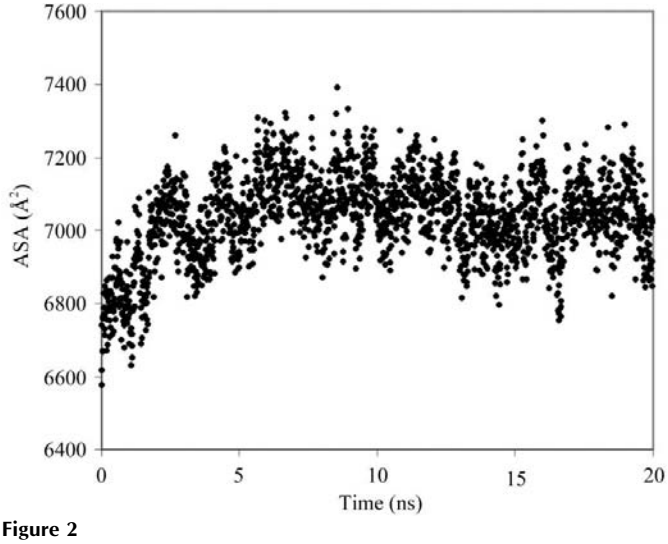

Figure 2
Fluctuations of the ASA values calculated for lysozyme during a 20 ns molecular-dynamics simulation. A possible explanation for the increase of the ASA values during the first 5 ns of the simulation is discussed in the text.

gation (applied to the formula for BSA above) shows that the variance (squared standard deviation) of the buried surface area is equal to the sum of the variances of the three ASA values. In other words, the error bar on the BSA value will exceed that on the total ASA value of the complex.

Taken together, the results presented here indicate that the effects of small random errors on calculated ASA values are of the order of 10–100 Å^2. For one small protein (HEWL), systematic effects arising from structural heterogeneity and possible model errors are of the order of 100 Å^2 and so are fluctuations arising from (simulated) dynamic effects. The error introduced by even one missing surface side chain can easily be 100 Å^2 and the spread observed when different programs are used to calculate ASA values for one small protein is almost 600 Å^2. On the positive side, for 69 HEWL structures the calculated ASA values were not significantly correlated with resolution or overall temperature factor. Moreover, buried surface areas of complexes can be reported with higher precision than the total ASA values of the individual proteins or the complex, provided that the calculations use only the crystal structure of the complex.

It is difficult to suggest what precision is warranted for reporting ASA values in general. The results presented here provide an impression of the size of the error bars and the factors that need to be taken into account. For total ASA values, a precision better than 10 Å^2 seems unrealistic in most cases and a value of 50–100 Å^2 prudent. It is hoped that the present findings will stimulate structural and computational biologists and bioinformaticians to give some thought to the issue of precision (and accuracy) of all properties that are calculated or derived from experimental structures.

The authors wish to thank Dr David van der Spoel (Uppsala) for stimulating discussions, Alexandra Patriksson (Uppsala) for providing the molecular-dynamics trajectory of lysozyme and Dr Roman Laskowski (EBI, Hinxton) for help with the *NACCESS* calculations. We also thank the editor and two anonymous referees for useful suggestions. MN is supported through grants to GJK from the Linnaeus Centre for Bioinformatics (Uppsala) and Uppsala University. MS was supported by a PhD fellowship from the Sven and Lilly Lawskis Foundation. GJK is a Royal Swedish Academy of Sciences (KVA) Research Fellow, supported through the Knut and Alice Wallenberg Foundation.

References

Artymiuk, P. J., Blake, C. C. F., Rice, D. W. & Wilson, K. S. (1982). *Acta Cryst.* B**38**, 778–783.
Bartels, C., Widmer, A. & Ehrhardt, C. (2005). *J. Comput. Chem.* **26**, 1294–1305.
Berendsen, H. J. C., Postma, J. P. M., DiNola, A. & Haak, J. R. J. (1984). *J. Chem. Phys.* **81**, 3684–3690.
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.
Cavallo, L., Kleinjung, J. & Fraternali, F. (2003). *Nucleic Acids Res.* **31**, 3364–3366.
Cole, C. & Warwicker, J. (2002). *Protein Sci.* **11**, 2860–2870.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cruickshank, D. W. (1999). *Acta Cryst.* D**55**, 583–601.
Darden, T., York, D. & Pedersen, L. (1993). *J. Chem. Phys.* **98**, 10089–10092.
Das, B. & Meirovitch, H. (2003). *Proteins*, **51**, 470–483.
Duan, Y., Reddy, B. V. & Kaznessis, Y. N. (2005). *Protein Sci.* **14**, 316–328.
Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. (1995). *J. Chem. Phys.* **103**, 8577–8592.
Gill, H. S., Pfluegl, G. M. & Eisenberg, D. (2002). *Biochemistry*, **41**, 9863–9872.
Henrick, K. & Thornton, J. M. (1998). *Trends Biochem. Sci.* **23**, 358–361.
Hubbard, S. J. & Thornton, J. M. (1993). *NACCESS.* Department of Biochemistry and Molecular Biology, University College London.
Inoue, T., Sugawara, H., Hamanaka, S., Tsukui, H., Suzuki, E., Kohzuma, T. & Kai, Y. (1999). *Biochemistry*, **38**, 6063–6069.
Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.
Kaliannan, P., Gromiha, M. M., Ramamurthi, K. & Elanthiraiyan, M. (1998). *Biophys. Chem.* **74**, 11–22.
Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2001). *J. Phys. Chem. B*, **105**, 6474–6487.
Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 842–857.
Kleywegt, G. J. (1997). *J. Mol. Biol.* **273**, 371–376.
Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.
Kleywegt, G. J., Harris, M. R., Zou, J.-Y., Taylor, T. C., Wahlby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.
Lee, B. & Richards, F. M. (1971). *J. Mol. Biol.* **55**, 379–400.
Mazumder-Shivakumar, D. & Bruice, T. C. (2005). *Biochemistry*, **44**, 7805–7817.
Nadassy, K., Wodak, S. J. & Janin, J. (1999). *Biochemistry*, **38**, 1999–2017.
Ondráček, J. & Mesters, J. R. (2006). *Acta Cryst.* D**62**, 996–1001.
Ondráček, J., Weiss, M. S., Brynda, J., Fiala, J., Jursík, F., Řezáčová, P., Jenner, L. B. & Sedláček, J. (2005). *Acta Cryst.* D**61**, 1181–1189.
Patriksson, A., Marklund, E. & van der Spoel, D. (2007). Submitted.
Rader, S. D. & Agard, D. A. (1997). *Protein Sci.* **6**, 1375–1386.
Richmond, T. J. (1984). *J. Mol. Biol.* **178**, 63–89.
Sanner, M. F., Olson, A. J. & Spehner, J. C. (1996). *Biopolymers*, **38**, 305–320.
Spoel, D. van der, Lindahl, E., Hess, B., Groenhof, G., Mark, A. E. & Berendsen, H. J. C. (2005). *J. Comput. Chem.* **26**, 1701–1718.
Willard, L., Ranjan, A., Zhang, H., Monzavi, H., Boyko, R. F., Sykes, B. D. & Wishart, D. S. (2003). *Nucleic Acids Res.* **31**, 3316–3319.