Molecular modelling of the mass density of single proteins

Meike Hutt, Tobias Kulschewski and Jürgen Pleiss*

Institute of Technical Biochemistry, Universität Stuttgart, Allmandring 31, D-70569 Stuttgart, Germany

Communicated by Ramaswamy H. Sarma

(Received 7 December 2011; final version received 27 February 2012)

Using molecular dynamics (MD) simulations, the density of single proteins and its temperature dependence was modelled starting from the experimentally determined protein structure and a generic, transferable force field, without the need of prior parameterization. Although all proteins consist of the same 20 amino acids, their density in aqueous solution varies up to 10% and the thermal expansion coefficient up to twofold. To model the protein density, systematic MD simulations were carried out for 10 proteins with a broad range of densities (1.32–1.43 g/cm³) and molecular weights (7–97 kDa). The simulated densities deviated by less than 1.4% from their experimental values that were available for four proteins. Further analyses of protein density showed that it can be essentially described as a consequence of amino acid composition. For five proteins, the density was simulated at different temperatures. The simulated thermal expansion coefficients ranged between 4.3 and 7.1 \times 10^{-4} \, \text{K}^{-1} and were similar to the experimentally determined values of ribonuclease-A and lysozyme (deviations of 2.4 and 14.6%, respectively). Further analyses indicated that the thermal expansion coefficient is linked to the temperature dependence of atomic fluctuations: proteins with a high thermal expansion coefficient show a low increase in flexibility at increasing temperature. A low increase in atomic fluctuations with temperature has been previously described as a possible mechanism of thermostability. Thus, a high thermal expansion coefficient might contribute to protein thermostability.

Keywords: molecular dynamics (MD) simulation; thermal expansion coefficient; amino acid composition; flexibility; thermostability

Introduction

Proteins are versatile molecules with interesting material properties. Based on sequence information only, they fold into a well-defined structure with high mechanical (Gosline, Demont, & Denny, 1986) and thermal stability (Singleton & Amelunxen, 1973), and have an amazing degree of flexibility (Huber, 1979). A delicate balance between structure and flexibility is essential for their biological function. All these material properties are coded in a linear sequence of a few 100 amino acids. Since the first description of sequence (Sanger & Tuppy, 1951a, 1951b) and structure (Kendrew et al., 1958) of proteins, understanding the properties of proteins on the basis of their sequence and structure is of utmost interest to biochemists and biotechnologists.

The density of a single protein (protein mass divided by protein volume) is a material property that since long has attracted attention, not only because it can be easily and precisely measured, but also because protein density and packing of amino acid side chains are the basis of many biophysical and biochemical properties such as stability towards heat (Goldstein, 2007) or organic solvents (Kawata & Ogino, 2010), flexibility and dynamics (Haliloglu, Bahar, & Erman, 1997; Halle, 2002) and aggregation (Meersman & Dobson, 2006). The densities of many proteins have been experimentally determined in dry state and solvated in water. At 20–25 °C, the majority of proteins have a density of about 1.35 g/cm³ in aqueous solution (Fischer, Polikarpov, & Craievich, 2004) and slightly less in dry state (Berlin & Pallansch, 1968; Brill, Siegel, & Olin, 1962), but some proteins have been identified with densities that are increased by almost up to 10% (Squire & Himmel, 1979).

To understand the molecular basis of the observed dependency of density on sequence and structure, different models have been suggested. For proteins with experimentally determined structure (Berman et al., 2000), the density has been evaluated from their volume. However, the absolute values of protein density
evaluated by different methods might differ considerably (Andersson & Hovmoller, 1998; Quillin & Matthews, 2000; Tsai, Taylor, Chothia, & Gerstein, 1999). For example, the protein density values of pyruvate kinase determined by different volume calculation methods ranged from 1.18 g/cm³ (Andersson & Hovmoller, 1998) to 1.47 g/cm³ (Quillin & Matthews, 2000). In a second approach, the densities of proteins were related to their molecular weight. By systematically comparing the densities of proteins with different molecular weights, it was observed that the densities are similar for all proteins with a molecular weight above 20 kDa (1.35 g/cm³) and increase gradually for low-molecular weight proteins (Fischer et al., 2004). Therefore, a molecular-weight-dependent function with an exponentially increasing protein density for decreasing molecular weights was proposed. However, this function cannot explain the observed large variations of density for proteins with similar low molecular weight. In a third approach, the density of proteins was correlated with their sequence and the amino acid composition of proteins was analysed. While for proteins in their dry state, the protein volumes could not be calculated from the sum of the volumes of its amino acids (Berlin & Pallansch, 1968), the experimentally determined density of proteins in solution could be well reproduced from their amino acid composition (Cohn & Edsall, 1943; Iqbal & Verrall, 1988; Kharakoz, 1997; Makhatadze, Medvedkin, & Privalov, 1990; Zamyatnin, 1972, 1984). While the densities of folded proteins vary by less than 10%, their experimentally determined thermal expansion coefficients vary by a factor of almost two (Chalikian, Totrov, Abagyan, & Breslauer, 1996). To account for thermal effects, the temperature dependence of the density of amino acid residues was determined (Amend & Helgeson, 2000; Hedwig & Hinz, 2003; Lee, Tikhomirova, Shalvardjian, & Chalikian, 2008; Makhatadze et al., 1990) and combined to model the temperature dependence of protein densities. However, the reproducibility of the experimental values using this method is limited as this strategy is primarily designed for unfolded proteins (Hedwig & Hinz, 2003). Furthermore, some approximations are made for the calculations. Therefore, the values determined over a wide temperature range are less reliable than those calculated for a temperature of 25°C (Hackel, Hinz, & Hedwig, 1999). Thus, a predictive model of protein density should describe properly the dependency of density not only on sequence and structure of the protein, but also on temperature. Therefore, we have developed a general molecular model to predict the density of a protein and its temperature dependence by using molecular dynamics (MD) simulations of protein-solvent systems. The method uses only the experimentally determined structure of a protein and a generic, transferable force field to evaluate the absolute value of protein density, and no prior parameterization is needed.

### Materials and methods

#### Setup of the simulations

Ten proteins were selected which cover a wide range of molecular weights from 7 to 97 kDa (Table 1). For the

| Protein | PDB structure | Molecular weight (kDa) | \( \rho_{\text{exp}} \) (g/cm³) | \( \rho_{\text{sim}} \) (g/cm³) | \( \rho_{\text{aa}} \) (g/cm³) |
|---------|---------------|------------------------|-----------------|-----------------|-----------------|
| 1 Scorpion toxin variant-3 | 2SN3 | 7.088 | – | 1.429 | 1.40 |
| 2 Human ubiquitin | 1UBI | 8.565 | – | 1.362 | 1.34 |
| 3 Bovine ribonuclease-A | 1XPS | 13.690 | 1.422 (Richard and Wyckoff 1971) | 1.405 | 1.41 |
| 4 Hen egg white lysozyme | 193L | 14.390 | 1.40 (Gekko and Noguchi 1979) | 1.381 | 1.39 |
| 5 Porcine adenylate kinase | 3ADK | 21.639 | 1.35 (Schirmer et al., 1970) | 1.366 | 1.34 |
| 6 Pseudomonas glumae lipase | 1TAH | 33.021 | – | 1.333 | 1.37 |
| 7 Candida antarctica lipase B | N74S mutant of 1TCA (Larsen, Bornscheuer, and Hult 2008) | 33.300 | – | 1.326 | 1.36 |
| 8 Guinea pig RP2 lipase | 1GPL | 47.788 | – | 1.362 | 1.37 |
| 9 Porcine malate dehydrogenase | 5MDH | 72.646 | 1.348 Gerding and Wolfe (1969) | 1.341 | 1.34 |
| 10 Rabbit glycogen phosphorylase B | 1GPB | 96.820 | – | 1.326 | 1.35 |

1With crystal water.
2Without crystal water.
3For one Candida antarctica lipase B molecule/box.
4For eight Candida antarctica lipase B molecules/box.
5For the monomer.
6For the dimer.
setup of the simulations, crystal water molecules and
ligands were removed from the protein structure. For
ubiquitin, two series of simulations with and without crys-
tal water were performed to estimate the effect of crystal
water on the simulated protein density. Subsequently, each
protein was solvated in four water boxes of different size,
according to protein concentrations of 8, 16, 24 and
32 mg/cm$^3$. The box sizes were in a range of 365–
20,062 nm$^3$ depending on the molecular weight of the
simulated protein and the protein concentration. For Can-
dida antarctica lipase B, a series of simulations was addi-
tionally performed with one and eight proteins in a box at
concentrations of 8, 12, 16, 20, 24, 28 and 32 mg/cm$^3$.
Moreover, the density of pure water was determined by
simulation of a large box of 46,657 water molecules.

The negatively charged proteins at pH 7 (Pseudomo-
nas glumae lipase (−1), Candida antarctica lipase B
(−1), RP2 lipase (−1), malate dehydrogenase (−4), gly-
cogen phosphorylase B (−7)) were neutralized by proton-
tating the respective number of histidines (the catalytic
histidine or histidines at the protein surface). The posi-
tively charged proteins (ribonuclease-A (+4), lysozyme
(+8), adenylate kinase (+2)) were neutralized by adding
chloride ions. For these three proteins, control simulations
were carried out with water boxes of corresponding size
containing water and the respective number of ions.

MD simulations were performed using the software
package groningen machine for chemical simulation
(GROMACS), version 4.0.7 (Van Der Spoel et al.,
2005). The simulations were performed at 298 K using the
optimized potentials for liquid simulations (OPLS)
all-atom force field (Jorgensen & Tirado-Rives, 1988)
and the extended simple point charge (SPC/E) water
model (Berendsen, Grigera, & Straatsma, 1987). For
ribonuclease-A and adenylate kinase, the simulations
were performed at 293 K, the temperature at which the
density was determined experimentally.

For an analysis of the temperature dependence of the
protein density, additional simulations of scorpion toxin
variant-3, ribonuclease-A, lysozyme, P. glumae lipase and
C. antarctica lipase B were done at 288 K and at 308 K.

Energy minimization, equilibration and production
phase
The systems were minimized using the steepest descent
algorithm with a maximum number of steps of 100,000.
Subsequently, all systems were equilibrated. The equili-
bration was started at an initial temperature of 10 K.
Within 200 ps, the systems were heated up to 100 K, fol-
lowed by 100 ps simulation at constant temperature.
Within the following 450 ps, the temperature was
increased to 288, 293, 298 or 308 K and equilibrated for
another 250 ps at the respective temperature. Temperature
coupling was done using velocity rescaling with a sto-
chastic term with a temperature coupling time constant
of .1 ps. The pressure was kept constant at 1 bar with a
pressure coupling time constant of .5 ps (NPT-model). A
time step of 1 fs was used.

After equilibration, the simulations were continued
for 10 ns with a time step of 2 fs. Temperature coupling
was performed using the Berendsen thermostat (Berends-
ken, Postma, van Gunsteren, DiNola, & Haak, 1984)
with a temperature coupling time constant of .1 ps and a
reference temperature of 298 K. Pressure coupling was
done using the Berendsen barostat (Berendsen et al.,
1984) with a pressure coupling time constant of 2 ps and
a constant pressure of 1 bar (NPT-model). Simulation of
10 ns was sufficient to allow for complete relaxation of
the solution density, which was constant during the last
5 ns of the simulation.

Sensitivity to simulation parameters
Four additional simulations of lysozyme were performed
with a different thermostat, barostat, water model or pro-
tein force field: (1) temperature coupling by the velocity-
rescaling thermostat (Bussi, Donadio, & Parrinello,
2007) was used instead of the Berendsen thermostat; (2)
pressure coupling by the Parrinello–Rahman barostat
(Parrinello & Rahman, 1981) was used instead of the
Berendsen barostat; (3) the water model TIP4P
(Jorgensen & Madura, 1985) was used instead of the
water model SPC/E; (4) the protein force field GROMOS96
43A1 (Christen et al., 2005) was used instead of the
OPLS all-atom force field.

Evaluation of protein density by MD simulation
The density of the protein solution $\rho_{\text{total}}$ consisting of
protein with mass $m_{\text{prot}}$ and volume $V_{\text{prot}}$ and water with
mass $m_{\text{wat}}$ and volume $V_{\text{wat}}$ is calculated by

$$\rho_{\text{total}} = \frac{m_{\text{prot}} + m_{\text{wat}}}{V_{\text{total}}} = c_{\text{prot}} + \frac{m_{\text{wat}}}{V_{\text{total}}} \quad (1)$$

With protein concentration $c_{\text{prot}} = \frac{m_{\text{prot}}}{V_{\text{total}}}$, box volume
$V_{\text{total}} = V_{\text{wat}} + V_{\text{wat}}$, water density $\rho_{\text{wat}} = \frac{m_{\text{wat}}}{V_{\text{wat}}}$ and protein
density $\rho_{\text{prot}} = \frac{m_{\text{wat}}}{V_{\text{wat}}}$, $m_{\text{wat}}$ can be expressed as

$$m_{\text{wat}} = \rho_{\text{wat}} \cdot V_{\text{wat}} = \rho_{\text{wat}} \left( V_{\text{total}} - \frac{m_{\text{prot}}}{\rho_{\text{prot}}} \right)$$

$$= V_{\text{total}} \cdot \rho_{\text{wat}} \left( 1 - \frac{c_{\text{prot}}}{\rho_{\text{prot}}} \right) \quad (2)$$

By inserting Equation (2) into Equation (1), the
solution density $\rho_{\text{total}}$ depends linearly on the protein
concentration $c_{\text{prot}}$ (Kupke, 1973):

$$\rho_{\text{total}} = \rho_{\text{wat}} + c_{\text{prot}} \left( 1 - \frac{\rho_{\text{wat}}}{\rho_{\text{prot}}} \right) \quad (3)$$
For each protein, simulations were performed for at least four different protein concentrations $c_{prot}$. For each simulation, the box volume was averaged over the last 5 ns of the simulation using the GROMACS tool g_energy, and the solution density $\rho_{total}$, the protein concentration $c_{prot}$ and their standard deviations were calculated. By plotting the densities of pure water and protein-water systems for at least four protein concentrations against the protein concentration, the protein den-
density $\rho_{prot}$ was calculated from the slope, and the water density $\rho_{wat}$ from the intercept with the $\gamma$-axis.

Thus, the calculation of protein density does not require an evaluation of protein surface or volume, and reflects amino acid packing and thermal motion.

**Evaluation of protein density by amino acid composition**

As an alternative method, the density was calculated based on the amino acid composition. For the calculation of the protein densities by amino acid composition, an additive scheme for the apparent specific volume of pro-
teins according to Cohn and Edsall (1943) was used (Table S1).

**Calculation of thermal expansion coefficients**

For the analysis of the temperature dependence of the pro-
tein density, protein density values were calculated at 288,
298 and 308 K, plotted against the corresponding temperature,
fitting by a straight line, and the thermal expansion coeffi-
cient was defined as the negative ratio of the slope
of the straight line to the protein density at 298 K.

**Protein structure and flexibility**

For analysis of protein structure, secondary structure elements were assigned according to dictionary of sec-
ondary structure of proteins (Kabsch & Sander, 1983). To analyse the flexibility of the simulated proteins, the root mean square fluctuation (RMSF) was calculated over the last 3 ns for all protein atoms using the GRO-
MACS tool g_rmsf. For each protein atom, the RMSF values were converted in the corresponding temperature
factor $B$ using the equation

$$B = \frac{8}{3} \cdot \pi^2 \cdot \text{RMSF}^2$$  \hspace{1cm} (4)

For each simulation, the median of the B-factor values of all protein atoms was calculated, and the median values were averaged over the four simulations at different protein concentrations. To analyse the variation of flexibility with temperature, median B-factors were calculated at 288, 298 and 308 K and plotted against the corresponding temperature. The correlation was fitted by a straight line, and the slope of the straight line was used as a measure of the increase in protein flexibility with temperature.

**Results**

**Quality of the protein simulations**

To confirm the quality of the water model for different temperatures, the density of a water box consisting of
46,657 water molecules at 298 K was determined by MD simulations. The calculations resulted in a value of .9960
$\pm .0005$ g/cm$^3$, which deviates by only .1% from the exper-
imental value (.9970 g/cm$^3$; Lide, 2009). Also at temperatures of 288 and 308 K, the simulations led to density
values that differed by only .1 and .3% from the experimental values, respectively. Thus, the SPC/E parameters are sufficiently accurate, and the system size and simulation time are large enough to reproduce experimental data.

To evaluate the quality of the protein simulations, the density of the 10 protein–water systems was followed as a function of simulation time. For all concentrations of all proteins, the density quickly relaxed to its equilibrium value during the first nanosecond of the simulation (Figure S1). Reproducible values with standard deviations of less than .1% could be obtained by averaging the solution density over a time interval of 5 ns. Thus, a simulation time of 10 ns was sufficient to determine the solution density. Furthermore, one simulation (the simu-
lation of eight proteins of Candida antarctica lipase B at a concentration of 20 mg/cm$^3$) was carried out for 20 ns instead of only 10 ns. Calculation of the solution density of this box over the time range of 5–10 and 15–20 ns led to identical values.

To evaluate the effect of simulating multiple proteins in a box as compared to a single protein, two series of simulations were carried out for C. antarctica lipase B, a single protein in a box and a system of eight proteins in a box (Figure S2). The protein density calculated by the two series deviated by .07% (Table 1). Thus, simulating a single protein in a box was sufficiently accurate, and protein–protein interactions in a realistic protein solution had no influence to the density. However, the observed relative standard deviations of the box volumes with one protein are slightly increased due to the smaller number of particles in the small boxes.

To evaluate whether variations in the density occur between the monomeric and dimeric form of a protein, simulations of malate dehydrogenase were done for the monomer as well as for the dimer (Figure S8). The resulting densities of the monomeric and dimeric malate dehydrogenase differed by .04% (Table 1). Therefore monomers and dimers of the same protein do not show different protein densities.

To evaluate the number of simulations at different protein concentrations that are necessary to reliably define a straight line, the simulations with C. antarctica lipase B were done for three additional protein concentrations. The resulting data points lay exactly on the straight line defined by four simulations at different concentration.
Four proteins (density of the 10 proteins was determined (Table 1). From the slope of these graphs, the corresponding to the temperature of the experimental density measurement. From the slope of these graphs, the corresponding to the temperature of the experimental density measurement. Thus, it was concluded that four protein concentrations plus the value for pure water were sufficient.

To evaluate the effect of crystal water to the evaluation of protein density, two series of simulations of ubiquitin (with and without crystal water) were compared (Figure S3). For both series, the protein densities differed by .1% (Table 1), thus the absence or presence of crystal water molecules does not influence the equilibration time of the system nor the resulting protein density.

For three positively charged proteins, chloride counterions were added to neutralize the protein system. To examine the influence of the ions on the solution density, boxes of corresponding sizes containing water and ions but without protein were simulated (Figure S5). The effect was analysed for lysozyme, for which eight chloride ions were added. Calculations of the protein density with and without consideration of the effect of ions on the solution density led to a protein density of 1.3809 and 1.4029 g/cm³, respectively. Thus, the influence of counterions should not be neglected upon calculation of the protein density.

To study whether the modelled density is sensitive towards changes in the thermostat, barostat, water model or force field, control simulations of lysozyme were performed. Replacing the SPC/E water with TIP4P water increased the simulated density by .17% (Table SII). Using the GROMOS force field instead of the OPLS all-atom force field increased the simulated density by .41%. Changes in the temperature coupling and pressure coupling decreased the simulated density by 1.2%. Using the Berendsen thermostat and barostat, the SPC/E water model and the protein OPLS all-atom force field, the simulated density of lysozyme deviated by −1.4% from the experimental value. Thus, when the v-rescale thermostat or the Parrinello–Rahman barostat were used, the experimental density was underestimated by 2.6%, with the TIP4P water model by 1.2%, and with the GROMOS force field by .95%.

Modelling the protein density by MD simulations

The densities of aqueous solutions of 10 proteins with molecular weights ranging from 7 to 97 kDa were determined by simulations of protein-water systems at four protein concentrations, 8, 16, 24 and 32 mg/cm³ for each protein (Figure 1 and Figures S2–S8). The simulations were performed at 293 or 298 K, corresponding to the temperature of the experimental density measurement. From the slope of these graphs, the density of the 10 proteins was determined (Table 1). Four proteins (C. antarctica lipase B, glycogen phosphorylase B, P. glumae lipase and malate dehydrogenase) had a low protein density (1.3257, 1.3259, 1.3330 and 1.3418 g/cm³, respectively), slightly lower than the average (1.35 g/cm³) of the majority of proteins. Four proteins (ubiquitin, RP2 lipase, adenylate kinase and lysozyme) had a medium protein density (1.3603, 1.3623, 1.3664 and 1.3809 g/cm³, respectively). Two proteins (ribonuclease-A and scorpion toxin variant-3) had a high density (1.4050 and 1.4290 g/cm³, respectively). As described previously (Fischer et al., 2004), the simulated density of a protein is related to its molecular weight: the molecular weight of the four proteins with low density is in the range 33–97 kDa, of the four proteins with medium density in the range 8.6–48 kDa, and of the two proteins with high density in the range 7–14 kDa (Figure 2). However, large differences in the protein densities are observed for the two small proteins, scorpion toxin variant-3 and ubiquitin.
For four proteins, experimental data of protein density in solution have been published previously. The simulated densities were in excellent agreement with their experimental values: for ribonuclease-A, the simulated density deviated by \(-1.2\)% from the experimental value (1.422 g/cm\(^3\) at 20°C [Richards 1971]); for lysozyme, the simulated density deviated by \(-1.4\)% from the experimental value (1.40 g/cm\(^3\) at 25°C) determined using a precision density meter (Gekko & Noguchi, 1979). This deviation is considerably smaller than a previous evaluation of the density of lysozyme (1.43 g/cm\(^3\)) (Tsai et al., 1999), corresponding to a deviation of \(+2.1\)% from the experimental value. For adenylate kinase, the simulated density was slightly larger (+1.2%) than the experimental value (1.35 g/cm\(^3\)) as determined by pycnometry (Schirmer, Schirmer, Schulz, & Thuma, 1970). For malate dehydrogenase, the simulated and the experimental density (1.348 g/cm\(^3\)) at 25°C (Gerding & Wolfe, 1969) deviated by less than 5.5%.

**Evaluation of protein density by amino acid composition**

To investigate the molecular reason for the variation in protein density by about 10%, the amino acid composition of the proteins was analysed, and the density was calculated as the sum of the experimentally determined specific volumes of the amino acids (Table SI) (Cohn & Edsall, 1943). According to the differences in density of the amino acids, proteins with a high content of hydrophobic residues have a lower density than proteins with a high content of hydrophilic residues (Figure 3). These calculations resulted in values for the protein densities that differed by only 7 and 1.3% on average from the values determined experimentally and by MD simulations, respectively (Table 1). Thus, the amino acid composition is the dominant factor that describes the protein density.

![Figure 3](image.png)

*Figure 3. Plot of the protein density against the content of hydrophobic amino acids (numbering of the proteins according to Table 1).*

---

**Table 2. Thermal expansion coefficients of scorpion toxin variant-3, ribonuclease-A, lysozyme, P. glumae lipase, and C. antarctica lipase B**

| Protein                        | Simulated thermal expansion coefficient \((10^{-4}K^{-1})\) |
|--------------------------------|----------------------------------------------------------|
| Scorpion toxin variant-3       | 7.1                                                      |
| Bovine ribonuclease-A           | 4.3 (4.2)                                                |
| Hen egg white lysozyme          | 4.7 (4.1)                                                |
| P. glumae lipase                | 6.1                                                      |
| C. antarctica lipase B          | 5.8                                                      |

**Analysing thermal expansion of proteins**

To evaluate the temperature dependence of the simulated density, simulations of five proteins (scorpion toxin variant-3, ribonuclease-A, lysozyme, P. glumae lipase and C. antarctica lipase B) were performed at three different temperatures (288, 298 and 308 K). In a temperature range of 20°C, the density depends linearly on the temperature (the \(R^2\) values for the linear fit of protein density and temperature are .9788 for scorpion toxin variant-3, .8005 for ribonuclease-A, .8462 for lysozyme, .9968 for P. glumae lipase and .9891 for C. antarctica lipase B), and the thermal expansion coefficient of the four proteins varies nearly twofold between 4.3 and 7.1 \(\times 10^{-4}K^{-1}\) (Table 2). For ribonuclease-A and lysozyme, the two proteins for which the thermal expansion coefficient was determined experimentally (Chalikian et al., 1996), the simulated and experimental values (deviations of 2.4 and 14.6%, respectively) agree well.

**Discussion**

**Modelled protein density**

To assess the quality of evaluating the protein density by MD simulations, the effects of simulation time, multiple proteins in a box, monomeric and dimeric forms, number of protein concentrations, crystal water and counterions were analysed, and the simulated densities were compared to the experimental values. The analysis revealed that simulation of 10 ns at four different protein concentrations is sufficient for a reliable determination of protein density. The number of proteins in a box, monomeric or dimeric forms of a protein and crystal water had no effect on the simulation result, whereas the effect of counterions on the solution density has to be considered in the calculation of protein density. The results of the MD simulation are in good agreement with the experiments. For four of the 10 simulated proteins (ribonuclease-A, lysozyme, adenylate kinase and malate dehydrogenase), experimental data on protein density are available. These four proteins cover a wide range of molecular weights (14–73 kDa) and...
densities (1.348–1.422 g/cm³). For these four proteins, the deviations between the simulated and the experimentally determined densities were less than 1.4%, thus we consider the strategy to model protein density by simulation of protein solutions at different concentrations as a reliable method to model protein density at high accuracy.

Furthermore, modelling of solvated proteins by MD simulations is a highly generic strategy, because protein force fields are based on a small set of force field parameters which are assigned by amino acid type and are independent of the environment of the amino acid. Thus, the same set of force field parameters applies to all proteins and a high transferability is guaranteed. With the exception of an initial choice of a force field, no further parameterization is needed to model the physical properties of molecular systems such as density. Modelling of protein density by MD simulation was insensitive towards the choice of the protein force field and the water model (<.5% deviation) and depended only slightly on the coupling method (1.2% for a different thermostat or barostat). This is an advantage and in contrast to previously suggested methods to calculate the density. In widely used methods to determine the protein density, the protein volume is calculated by defining the protein surface. Using the GROMACS tool g_sas (van der Spoel et al., 2005), the density of hen egg white lysozyme varied between 1.05 and 1.43 g/cm³ upon changing the probe radius from 1.4 to 7 Å. Similarly, for the same set of representative proteins, two different algorithms resulted in different values for the average protein density: 1.43 and 1.47 g/cm³ as calculated by the Connolly and the Voronoi algorithm, respectively (Quillen & Matthews, 2000). These values deviate by 6 and 9%, respectively, from the average protein density of 1.35 g/cm³ determined experimentally (Fischer et al., 2004).

Molecular reasons for the variation of protein density

While all proteins consist essentially of the same 20 amino acids, their densities deviate by up to almost 10%. It was observed that proteins with a molecular weight of less than 20 kDa tend to have an increased density (Fischer et al., 2004). This tendency was essentially confirmed by analysing experimentally determined protein densities (Squire & Himmel, 1979), values calculated by Tsai et al. (1999) and by MD simulations as a function of molecular weight (Figure 2). However, for proteins of similar molecular weight, large deviations of the protein density may occur. The two proteins, scorpion toxin variant-3 and ubiquitin, have a similar molecular weight below the 20 kDa threshold (7.1 and 8.6 kDa, respectively), but have significantly different simulated protein densities (1.43 and 1.36 g/cm³, respectively). The density of ubiquitin is similar to the considerably larger RP2 lipase (48 kDa). Therefore, a simple dependence of the protein density on the molecular weight is not adequate to describe the molecular basis of protein density.

To explain the observed differences in protein densities, it was suggested that the amino acid composition determines protein density (Cohn & Edsall, 1943; Iqbal & Verrall, 1988; Kharakoz, 1997; Makhadzade et al., 1990; Zamyatnin, 1972, 1984). Indeed, proteins with high density have a low percentage of hydrophobic amino acids (scorpion toxin variant-3: ρ = 1.4290 g/cm³, 17% hydrophobic residues), while proteins with low density have a high percentage of hydrophobic amino acids (ubiquitin: ρ = 1.3603 g/cm³, 33% hydrophobic residues, RP2 lipase: ρ = 1.3623 g/cm³, 32% hydrophobic residues). This observation is in accordance with the low density and high specific volume v of hydrophobic amino acids as compared to hydrophilic amino acids (Ile: v = 0.90 cm³/g, Asp: v = 0.60 cm³/g [Cohn & Edsall, 1943]). Thus, the observation that small proteins in general have higher densities than large proteins is a consequence of their low content of hydrophobic amino acids. As small proteins have higher surface to volume ratios than large proteins and as the surface is hydrophilic, small proteins have a higher content of hydrophilic amino acids and thus a higher density. Therefore, the density of a protein can be predicted from its sequence (Iqbal & Verrall, 1988; Kharakoz, 1997; Makhadzade et al., 1990; Zamyatnin, 1984), and the results obtained by MD simulations are in good agreement with the calculations from amino acid composition (maximum deviation of 2.7%).

Temperature dependence of protein density

Simulations of scorpion toxin variant-3, ribonuclease-A, lysozyme, P. glumae lipase and C. antarctica lipase B in a temperature range of 288–308 K were performed to determine the thermal expansion coefficients of these five proteins. The thermal expansion coefficients of ribonuclease-A and lysozyme determined by MD simulation are in good agreement with the experimental values (deviations of 2.4 and 14.6%, respectively), thus modelling of protein density by MD simulations is valid for a range of temperatures.

In contrast to protein densities which vary by less than 10%, the thermal expansion coefficients of the five proteins investigated here vary nearly twofold (4.3–7.1 × 10⁻⁴ K⁻¹). This high variation of the thermal expansion coefficients was confirmed by experimental data on proteins such as hen egg white lysozyme (4.1 × 10⁻⁴ K⁻¹) and bovine serum albumin (7.1 × 10⁻⁴ K⁻¹) (Chalikian et al., 1996). Considering the fact that proteins consist of the same material, the variation is surprisingly large. Therefore, it would be interesting to understand the molecular basis of the thermal expansion of proteins. Interestingly, there is no correlation between the thermal expansion of a protein and its...
protein density at 298 K (Figure S9). The two proteins with a low density (P. glumae lipase and C. antarctica lipase B) have a similar thermal expansion coefficient (6.1 and $5.8 \times 10^{-4} \text{K}^{-1}$, respectively), while the two proteins with high density (ribonuclease-A and scorpion toxin variant-3) have different thermal expansion coefficients (4.3 and $7.1 \times 10^{-4} \text{K}^{-1}$, respectively). Lysozyme has a medium density and a thermal expansion coefficient ($4.7 \times 10^{-4} \text{K}^{-1}$) similar to that of ribonuclease-A.

As another approach to analyse molecular determinants of thermal expansion, correlations between thermal expansion and secondary structure composition and its variation with temperature were analysed. However, no correlations of thermal expansion and secondary structure composition and its variation with temperature were analysed. However, no correlations of thermal expansion and secondary structure composition could be shown (Figure S10) and no change in secondary structure composition in the analysed temperature range was observed (Figure S11).

Because the thermal fluctuations increase with temperature, the correlation between protein flexibility and density at different temperatures was investigated. Therefore, the median B-factor at 298 K and the variation of protein flexibility with temperature were analysed. As the median equals to the maximum RMSF of the 50% most rigid protein atoms, it measures the flexibility of the core of the protein and therefore, in contrast to the average value, is independent of individual, highly flexible loop regions. Intuitively, we expected a positive correlation between the thermal expansion coefficient and the temperature dependence of the protein flexibility. In contrast, we found a negative correlation between thermal expansion coefficient and flexibility, and between thermal expansion coefficient and temperature dependence of protein flexibility: proteins with a high thermal expansion coefficient have a low median B-factor at 298 K (Figure S12) and a low increase in protein flexibility with temperature (Figure 4). This surprising observation could be explained by the difference between RMSF and configurational entropy (Missimer et al., 2007). It has been shown that the configurational entropy increases with temperature (Missimer et al., 2007), and the configurational entropy of glasses can be expressed in terms of molar thermal expansion (Casalini, Capaccioli, Lucci, Rolla, & Coreazzi, 2001). However, the RMSF does not strictly follow the configurational entropy: while the RMSF increased strongly with temperature for a thermolabile protein, it increased less for a more thermostable protein complex, indicating that the intra-molecular interactions in thermostable proteins restrict the increase in fluctuations at increasing temperatures despite an increase in configurational entropy (Missimer et al., 2007). A broad variation in the temperature dependence of the protein flexibility was also observed for the five proteins investigated here: for two proteins (scorpion toxin variant-3 and C. antarctica lipase B), the protein flexibility varied only slightly with temperature (.07Å²/K), while the atomic fluctuations of ribonuclease-A increased strongly with temperature (.27Å²/K). Thus, a high thermal expansion coefficient results in a slow increase in atomic fluctuations at increasing temperature. Because thermal fluctuations decrease the kinetic barrier to unfolding, a high thermal expansion coefficient is expected to contribute to protein stabilization. Increasing the thermal expansion coefficient would then be an appropriate design strategy for creating thermostable proteins. However, it is not yet clear how the thermal expansion coefficient is encoded in the sequence and structure of a protein, but it would be highly interesting to study in more detail the molecular basis of protein expansion.

**Conclusion**

Using MD simulations, protein density and its temperature dependence were modelled based only on the sequence and structure of a protein and a generic, transferable force field without the need of prior parameterization. The method was applied to evaluate two basic material properties of proteins, density and thermal expansion, which were in good agreement with experimental values. The protein density is determined by the amino acid composition of the protein and is independent of molecular weight. The thermal expansion coefficient is linked to the temperature-dependent increase in protein flexibility: proteins with a high thermal expansion coefficient show a small increase in flexibility at increasing temperature, indicating that a high thermal expansion coefficient might contribute to protein thermostability.

**Acknowledgements**

We thank the High Performance Computing Center Stuttgart for providing access to their computer facilities, and the Deutsche Forschungsgemeinschaft (SFB716) for financial support.
Supplementary Material

The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2012.680031.

References

Amend, J.P., & Helgeson, H.C. (2000). Calculation of the standard molal thermodynamic properties of aqueous biomolecules at elevated temperatures and pressures II. Unfolded proteins. Biophysical Chemistry, 84(2), 105–136.

Andersson, K.M., & Hovmoller, S. (1998). The average atomic volume and density of proteins. Zeitschrift für Kristallographie [Journal of Crystallography], 213(7–8), 369–373.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., & Haak, J.R. (1984). Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics, 81(8), 3684–3690.

Berlin, E., & Pallansch, M.J. (1968). Densities of several proteins and L-amino acids in dry state. Journal of Physical Chemistry, 72(6), 1887–1899.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, N.T., Weissig, H., … Bourne, P.E. (2000). The protein data bank. Nucleic Acids Research, 28(1), 235–242.

Brill, A.S., Siegel, B.M., & Olin, J.S. (1962). Specific volume of dry protein. Journal of Cell Biology, 13(2), 249–251.

Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through velocity rescaling. The Journal of Chemical Physics, 126(1), 014101–014107.

Casalini, R., Capaccioli, S., Lucchesi, M., Rolla, P.A., & Corezzi, S. (2001). Pressure dependence of structural relaxation time in terms of the Adam-Gibbs model. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 63(3 Pt 1), 031207.

Chalikian, T.V., Totrov, M., Abagyan, R., & Breslauer, K.J. (1996). The hydration of globular proteins as derived from volume and compressibility measurements: Cross correlating thermodynamic and structural data. Journal of Molecular Biology, 260(4), 588–603.

Christen, M., Hünenberger, P.H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D.P., … van Gunsteren, W.F. (2005). The GROMOS software for biomolecular simulation: GROMOS96. Journal of Computational Chemistry, 26(16), 1719–1751.

Cohn, E.J., & Edsall, J.T. (1943). Density and apparent specific volume of proteins. In E.J. Cohn & J.T. Edsall (Eds.), Proteins, amino acids and peptides (pp. 370–381). New York, NY: Reinhold.

Fischer, H., Polikarpov, I., & Craievich, A.F. (2004). Average protein density is a molecular-weight-dependent function. Protein Science, 13(10), 2825–2828.

Gekko, K., & Noguchi, H. (1979). Compressibility of globular-proteins in water at 25°C. Journal of Physical Chemistry, 83(21), 2706–2714.

Gerdig, R.K., & Wolfe, R.G. (1969). Malic dehydrogenase 8. Large scale purification and properties of supernatant pig heart enzyme. Journal of Biological Chemistry, 244(5), 1164–1171.

Goldstein, R.A. (2007). Amino-acid interactions in psychrophiles, mesophiles, thermophiles, and hyperthermophiles: Insights from the quasi-chemical approximation. Protein Science, 16(9), 1887–1895.

Gosline, J.M., Demont, M.E., & Denny, M.W. (1986). The structure and properties of spider silk. Endeavour, 10(1), 37–43.

Hackel, M., Hinz, H.J., & Hedwig, G.R. (1999). Partial molar volumes of proteins: Amino acid side-chain contributions derived from the partial molar volumes of some tripeptides over the temperature range 10–90°C. Biophysical Chemistry, 82(1), 35–50.

Halligoulu, T., Bahar, I., & Erman, B. (1997). Gaussian dynamics of folded proteins. Physical Review Letters, 79(16), 3090–3093.

Halle, B. (2002). Flexibility and packing in proteins. Proceedings of the National Academy of Sciences of the United States of America, 99(5), 1274–1279.

Hedwig, G.R., & Hinz, H.J. (2003). Group additivity schemes for the calculation of the partial molar heat capacities and volumes of unfolded proteins in aqueous solution. Biophysical Chemistry, 100(1–3), 239–260.

Huber, R. (1979). Conformational flexibility and its functional significance in some protein molecules. Trends in Biochemical Sciences, 4(12), 271–276.

Iqbal, M., & Verrall, R.E. (1988). Implications of protein folding – additivity schemes for volumes and compressibilities. Journal of Biological Chemistry, 263(9), 4159–4165.

Jorgensen, W.L., & Madura, J.D. (1985). Temperature and size dependence for Monte Carlo simulations of TIP4P water. Molecular Physics, 56(6), 1381–1392.

Jorgensen, W.L., & Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crombin. Journal of the American Chemical Society, 110(6), 1657–1666.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22(12), 2577–2637.

Kawata, T., & Ogino, H. (2010). Amino acid residues involved in organic solvent-stability of the LST-03 lipase. Biochemical andBiophysical Research Communications, 400(3), 384–388.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., WyckoFF, H., & Phillips, D.C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature, 181(4610), 662–666.

Kharakoz, D.P. (1997). Partial volumes and compressibilities of extended polypeptide chains in aqueous solution: Additivity scheme and implication of protein unfolding at normal and high pressure. Biochemistry, 36(33), 10276–10285.

Kupke, D.W. (1973). Density and volume change measurements. In S.J. Leach (Ed.), Physical principles and techniques of protein chemistry, part C (pp. 1–75). New York, NY: Academic Press.

Larsen, M.W., Bornscheuer, U.T., & Hult, K. (2008). Expression of Candida antarctica lipase B in Pichia pastoris and various Escherichia coli systems. Protein Expression and Purification, 62(1), 90–97.

Lee, S., Tikhomirova, A., Shalvardjian, N., & Chalikian, T.V. (2008). Partial molar volumes and adiabatic compressibilities of unfolded protein states. Biophysical Chemistry, 134(3), 185–199.

Lide, D.R. (2009). CRC Handbook of chemistry and physics. Boca Raton, FL: Taylor & Francis.
Makhatadze, G.I., Medvedkin, V.N., & Privalov, P.L. (1990). Partial molar volumes of polypeptides and their constituent groups in aqueous solution over a broad temperature range. *Biopolymers, 30*(11–12), 1001–1010.

Meersman, F., & Dobson, C.M. (2006). Probing the pressure-temperature stability of amyloid fibrils provides new insights into their molecular properties. *Biochimica Et Biophysica Acta-Proteins and Proteomics, 1764*(3), 452–460.

Missimer, J.H., Steinmetz, M.O., Baron, R., Winkler, F.K., Kammerer, R.A., Daura, X., & van Gunsteren, W.F. (2007). Configurational entropy elucidates the role of salt-bridge networks in protein thermostability. *Protein Science, 16*(7), 1349–1359.

Parrinello, M., & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics, 52*(12), 7182–7190.

Quillin, M.L., & Matthews, B.W. (2000). Accurate calculation of the density of proteins. *Acta Crystallographica Section D-Biological Crystallography, 56*, 791–794.

Richards, F.M., & Wyckoff, H.W. (1971). Bovine pancreatic ribonuclease. In P.D. Boyer (Ed.), *The enzymes* (Vol. 4, pp. 647–806). New York, NY: Academic Press.

Sanger, F., & Tuppy, H. (1951a). The amino-acid sequence in the phenylalanyl chain of insulin 2. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal, 49*(4), 481–490.

Sanger, F., & Tuppy, H. (1951b). The amino-acid sequence in the phenylalanyl chain of insulin I. The identification of lower peptides from partial hydrolysates. *Biochemical Journal, 49*(4), 463–481.

Schirmer, L., Schirmer, R.H., Schulz, G., & Thuma, E. (1970). Purification, characterization and crystallization of pork myokinase. *FEBS Letters, 10*(5), 333–338.

Singleton, R., & Amelunxen, R.E. (1973). Proteins from thermophilic microorganisms. *Bacteriological Reviews, 37*(3), 320–342.

Squire, P.G., & Himmel, M.E. (1979). Hydrodynamics and protein hydration. *Archives of Biochemistry and Biophysics, 196*(1), 165–177.

Tsai, J., Taylor, R., Chothia, C., & Gerstein, M. (1999). The packing density in proteins: Standard radii and volumes. *Journal of Molecular Biology, 290*(1), 253–266.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., & Berendsen, H.J. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry, 26*(16), 1701–1718.

Zamyatnin, A.A. (1972). Protein volume in solution. *Progress in Biophysics and Molecular Biology, 24*, 107–123.

Zamyatnin, A.A. (1984). Amino-acid, peptide, and protein volume in solution. *Annual Review of Biophysics and Bioengineering, 13*, 145–165.