Self-assembling dipeptides: conformational sampling in solvent-free coarse-grained simulation

Alessandra Villa, Christine Peter and Nico F. A. van der Vegt*

We discuss the development of a coarse-grained (CG) model for molecular dynamics (MD) simulation of a hydrophobic dipeptide, diphenylalanine, in aqueous solution. The peptide backbone is described with two CG beads per amino acid, the side groups and charged end groups are each described with one CG bead. In the derivation of interaction functions between CG beads we follow a bottom-up strategy where we devise potentials such that the resulting CG simulation reproduces the conformational sampling and the intermolecular interactions observed in an atomistic simulation of the same peptide. In the CG model, conformational flexibility of the peptide is accounted for through a set of intra-molecular (bonded) potentials. The approach followed to obtain the bonded potentials is discussed in detail. The CG potentials for nonbonded interactions are based on potentials of mean force obtained by atomistic simulations in aqueous solution. Following this approach, solvent mediation effects are included in the effective bead–bead nonbonded interactions and computationally very efficient (solvent-free) simulations of self-assembly processes can be performed. We show that the conformational properties of the all-atom dipeptide in explicit solvent can be accurately reproduced with the CG model. Moreover, preliminary simulations of peptide self-assembly performed with the CG model illustrate good agreement with results obtained from all-atom, explicit solvent simulations.

1. Introduction

Coarse-grained (CG) models for biomolecular systems are used to study (among others) the folding of proteins, phase behavior of lipids and the formation of vesicles, bilayers, *etc*. These models play an important role in biological physics and have been particularly successful unravelling several biophysical phenomena at mesoscopic time and length scales.[1–6] Linking these types of models to more detailed particle models that include chemical specificity requires scale-bridging approaches. These approaches have come of age during the past decade and link detailed atomistic and coarse-grained models in such a way that the resolution of the model can be changed at will, either in sequential simulations or in a single (multiscale) simulation with adaptive resolution.[7,8] Although there have already been quite a number of applications in this direction, the current status of the field is such that we are faced with challenges in how to treat nonbonded interactions at the coarse-grained level (to model *e.g.* self-assembly in chemical specific systems) and how to treat intramolecular, bonded interactions of coarse-grained biomolecules (to model *e.g.* conformational changes).

Various approaches have been used to develop CG nonbonded interactions. One possibility is to assume a functional form of the nonbonded interaction between "superatoms" and parameterize the force field on thermodynamic data[9–12] (in a spirit similar to how atomistic force fields are parameterized), often following hydrophilicity–hydrophobicity arguments.[13–17] Another approach to deriving CG nonbonded interactions relies on the sampling of a detailed system and projecting a set of detailed forces onto a smaller set of coarse-grained forces.[18,19] The so called "structure-based" CG methods provide CG nonbonded interactions that reproduce a pre-defined target structure—often described by a set of radial distributions functions obtained from all-atom molecular simulations.[20–23] These latter methods do not make any assumption on the functional form of the nonbonded potential, which is obtained in tabulated form.

Also for covalent/bonded interactions, several CG approaches have been developed. Especially in the field of peptides and proteins, most parametrizations are either based on general conformational properties of the polypeptide backbone ("Ramachandran plots"),[13–15,24] or protein conformations are constrained to an experimental structure.[25,26] In polymer simulations, bonded interactions in CG models have received quite some attention. Here chain conformations sampled with coarse-grained models have to be consistent with detailed-atomistic realizations of chain conformations.[27–30] The methods used to derive CG bonded potentials for polymeric systems can at least in principle be carried over to biopolymers, but have not been explored so far. Obviously, the challenges in developing CG bonded potentials for biomolecules (without build-in biases that drive the system to sample certain states) are significantly greater due to the vast number of primary sequences, secondary and tertiary structures, *etc*. found for example in polypeptides and proteins as

Max Planck Institute for Polymer Research, Ackermannweg 10, D-55128 Mainz, Germany. E-mail: vdervegt@mpip-mainz.mpg.de; Fax: +49 (0)6131 379 340; Tel: +49 (0)6131 379 245

well as due to the nontrivial dependence of these structures on the chemical environment.

As a first step towards CG peptide models we present a CG model for a hydrophobic dipeptide in water. The modelling approach is based on an underlying atomistic description, *i.e.* it uses a bottom-up CG strategy aiming at a multiscale description that allows us to switch the level of resolution between the CG and the atomistic scale. We will discuss the development of bonded and nonbonded CG potentials and apply the model to study the self-assembly of dipeptides at a finite concentration in water. In this paper we outline in detail how we introduce conformational flexibility in the coarse-grained dipeptide model. This aspect is of particular importance because in many biological processes the conformations may depend on the immediate environment of the bio-molecule. We will also discuss the construction of nonbonded CG potentials, which, in this paper, will be based on an implicit-solvent representation. In the next paper in this issue, we extend the dipeptide model to include an explicit, yet coarse-grained, solvent representation. Finally, we will illustrate how inverse-mapping of chemical details can be performed to study the packing of individual peptides within self-assembled aggregates sampled with the CG model.

2. Coarse graining approach

The following section outlines the development of the CG model, with special focus on aspects concerning covalent/bonded interactions. The latter are particularly challenging in the case of biomolecules compared to many synthetic polymers due to aspects such as sequence information, chirality, *etc.* and consequently complex conformational equilibria.

2.1 Mapping scheme

The mapping scheme relates the bead positions in the CG model with the atomistic coordinates. In the present model the peptide, diphenylanaline, is described by seven spherically symmetric beads as shown in Fig. 1.

Five different bead types were identified (the position of the CG bead is constituted as the center of mass of the underlying atoms): an N-terminal bead denoted further on as "NH$_3$"; a bead type representing the α and β carbon atoms ("C$\alpha\beta$"); a
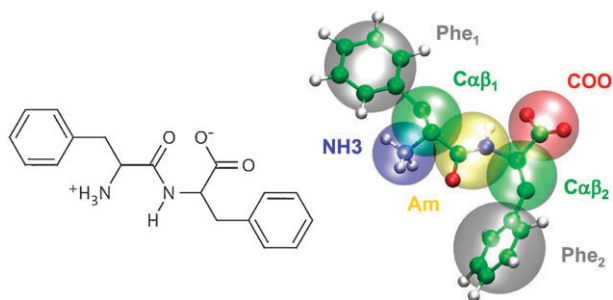

Fig. 1 Atomistic structure of diphenylalanine and mapping scheme. Aliphatic hydrogen atoms are omitted. The 7 CG beads are located on the centers-of-mass of the following groups of atoms: NH$_3$, C$_\alpha$–C$_\beta$, 6 phenyl aromatic carbons, C$_\alpha$–CO–NH–C$_\alpha$ and COO. The label indicates the bead type and residue number.

bead type representing the aromatic ring of the sidechains ("Phe"); a bead type representing the amide group ("Am", the mapped coordinates given by the center of mass of the four atoms in the peptide group and the two adjacent α carbon atoms); and a C-terminal bead ("COO").

2.2 Covalent/bonded interaction potentials

The parameterization of CG covalent interaction potentials aims at reproducing the conformational sampling of the atomistic system.

First, we performed extensive conformational sampling of a single peptide in water by atomistic simulations, obtained the corresponding CG coordinates according to the mapping scheme and determined the distributions of CG bond lengths, $P^{CG}(r,T)$, CG angles, $P^{CG}(\theta,T)$, and CG dihedral angles $P^{CG}(\phi,T)$. These distributions can be converted into potentials (potentials of mean force) through Boltzmann inversion:

$$U^{CG}(r,T) = -k_B T \ln(P^{CG}(r,T)/r^2) + C_r \quad (1)$$

$$U^{CG}(\theta,T) = -k_B T \ln(P^{CG}(\theta,T)/\sin(\theta)) + C_\theta \quad (2)$$

$$U^{CG}(\phi,T) = -k_B T \ln P^{CG}(\phi,T) + C_\phi . \quad (3)$$

k_B denotes the Boltzmann constant, T the temperature.

The thus derived potentials do not necessarily yield a CG model which reproduces the sampling of the atomistic system for several reasons:

The first reason is the assumption that covalent and non-bonded interactions are independent and can be parametrized separately. This assumption stems from the structure-based coarse graining methodology that had been originally developed for amorphous polymer systems. There, the distributions to determine bonded potentials are usually obtained by excluding (long-range) nonbonded interactions within the chain in the atomistic sampling (those that go beyond the distance of a CG torsion). The reason for this is that one tries to avoid double counting interactions. Those interactions that are taken care of in the CG simulations through nonbonded potentials between CG particles, are not included when the covalent potentials are being determined, while those CG beads that interact in principle through covalent potentials are excluded nonbonded-wise on the CG level. This clean separation assumes that the CG nonbonded "environment" has the same effect on the conformational sampling as the all-atom nonbonded "environment", which is well-justified in the case of amorphous polymers. This assumption needs to be assessed carefully in the case of biomolecules where hydrogen bonding, solvation effects, *etc.* drastically influence the con-formational sampling. In this study, we determined the cova-lent interaction potentials for the CG model of the dipeptide based on two sets of atomistic simulations: one where the atomistic peptide system was simulated with all nonbonded interactions present (set "noex"), and one, where we excluded nonbonded interactions between atoms that correspond to CG particles that are further than three bonds apart (correspond-ing to a CG torsion angle interaction) (set "ex"). The reason for this particular exclusion scheme is that in the CG simula-tions the interactions between these excluded pairs are taken care of by intramolecular nonbonded potentials. Practically

this means that, in the atomistic sampling runs, interactions between atoms "belonging" to the following bead pairs are excluded: COO–NH$_3$, COO–Phe$_1$, NH$_3$–Phe$_2$ and Phe$_1$–Phe$_2$. The set of parameters determined from the two sampling are labeled CG^{noex} and CG^{ex}, respectively. In principle this latter sampling should give "cleaner" potentials from the perspective of "double counting" of interactions, but it needs to be carefully tested whether the intramolecular nonbonded interactions between CG beads have the same effect on the conformational sampling as the corresponding atomistic interactions.

Second, the above potentials can only correctly reproduce the conformational sampling of the peptide in the atomistic description if all degrees of freedom are uncorrelated, *i.e.* if the probability distribution describing the conformations of the peptide cleanly factorizes into bond, angle and torsion contributions:

$$P^{CG}(r,\theta,\phi,T) = P^{CG}(r,T)P^{CG}(\theta,T)P^{CG}(\phi,T). \qquad (4)$$

Only if this is true, the distributions may be Boltzmann-inverted separately into potentials (eqn (1)–(3)) which can be applied additively. Let us assume two internal CG degrees of freedom that can be described by potential energy functions with two minima each. This results in four conformational states that could in principle be occupied. In the case where the CG degrees of freedom are correlated it is well possible that two of these parameter combinations correspond to "forbidden" conformations in the atomistic sampling but are "allowed" by the CG sampling. Another scenario is that the states are correlated and the CG sampling correctly captures that some parameter combinations are "forbidden" but by applying *two* potentials (one for each degree of freedom) that affect the same two (remaining) states, we have effectively applied a too high barrier between the two states.

This latter scenario is particularly likely to be the case if we need to apply several covalent interaction potentials that involve the same CG particles. One example is the CG bead that corresponds to the α and β carbon (C$\alpha\beta$ bead type) of the peptide backbone, where the sidechain bead branches off. This bead is involved in backbone torsions, three angles and one additional potential, the latter corresponding to a so called improper dihedral (taking care of out of plane motions) that is needed to preserve the chirality of the amino acid (see Results and discussion).

2.3 Nonbonded interaction potentials

In the CG bonded potentials discussed above, solvation effects have effectively been included by performing the atomistic sampling of the dipeptide conformations in explicit solvent. In the following, we will determine nonbonded interaction potentials for the CG peptide beads that reproduce the peptide–peptide interactions in the atomistic sampling resulting in an implicit solvent description.[31–33]

Because our CG model has five different bead types, a total number of 15 bead–bead nonbonded interaction potentials need to be constructed. We will assume that each of these bead–bead interaction potentials can be described with a pair potential of mean force (PMF) in which thermal averaging

over the bead internal degrees of freedom and the solvent degrees of freedom has been performed and the only remaining degree of freedom is the bead–bead radial distance. The so-obtained pair interaction potentials include the effects of solvent mediation, which are crucial in obtaining a realistic CG description with the appropriate balance of hydrophobic interactions between the nonpolar beads and the hydrophilic interactions between the polar and charged beads.

For each of the 15 nonbonded parameter combinations, an all-atom pair PMF was calculated using a single solute pair in explicit water with constrained solute–solute distances (see Methods/computational details). We used a different solute for each bead type (see Fig. 1). Methane, benzene and methylacetamide molecules model the C$\alpha\beta$, Phe and Am beads, respectively, while methylammonium and acetate ions model the peptide termini (NH$_3$ and COO). The solute centers used to define the distances in the constraint simulations correspond to the mapping points of the CG peptide beads. Fig. 2 shows as an example the resulting effective potentials for polar (amide–amide), ionic (acetate–acetate) and hydrophobic (benzene–benzene) beads in water solution. The thus obtained nonbonded potentials are used for intermolecular interactions between different peptides as well as intramolecular interactions between beads that are separated by more than three bonds (*i.e.* between Phe beads and peptide termini).

It should be noted here that the above approach of determining effective interaction functions between peptide beads based on small molecules, *i.e.* fragments of the target molecule, does not necessarily result in a set of potentials that reproduces the peptide–peptide interaction and peptide self-assembly as observed at the atomistic level. First, the small molecules might sample relative conformations or orientations that—due to bonded interactions and steric hindrance—are not relevant for the corresponding fragments within the chain, but these conformations might nevertheless significantly contribute to the CG peptide model. Second, the approach only holds if the fragment-based CG nonbonded potentials are additive when stringed together in a whole molecule. Third, the interactions were determined based on pair potentials of mean force, *i.e.* the effective interaction at infinite dilution, whereas the peptide concentration during self-assembly is considerably higher. These assumptions need to be validated, as we will discuss in more detail in the Results and discussion section.
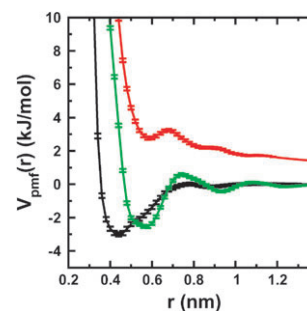

Fig. 2 Potential of mean force, $V_{PMF}(r)$, between two methylacetamide molecules (black), two acetate ions (red) and two benzene molecules (green) in water solution at 300 K.

2.4 Inverse mapping

Inverse mapping is the procedure to reintroduce atomistic details into the CG structure. It should be noted that every CG configuration corresponds to an average over many atomistic microstates. Thus, the inverse mapping procedure generates one microstate out of this set. Different strategies are available to reintroduce the atomistic details into the CG structure.[20,27,34,35] Here, the inverse mapping was performed in the following steps:

1. The initial atomistic coordinates for the peptide were randomly generated around the position of the corresponding CG bead.

2. Nonbonded interactions were switched off, while the bonds were made flexible and three extra bond interactions were added between the opposite phenyl carbon atoms to facilitate ring formation. The CG bead positions were related to the atomistic coordinate according to the mapping scheme using a virtual site. A short energy minimization (100 steps) was performed to obtain a proper molecular geometry. During the minimization, the atoms were forced by an additional harmonic (restraining) potential to match the location of the corresponding CG bead (virtual site).

3. The peptide was solvated and intra- and inter-molecular nonbonded interactions were switched on. A short energy minimization (100 steps) is used to relax the position of the water around the peptide while virtual site restraining was applied to peptides atoms. Standard bond potentials were used.

4. A short unrestrained simulations of 2 ps was performed to relax and equilibrate the atomistic peptide–water system.

3. Results and discussion

3.1 Comparison of atomistic and CG conformations

In order to assess the ability of the covalent CG potentials determined through Boltzmann inversion to reproduce the correct peptide conformations, we compare the conformational sampling obtained by atomistic and CG simulations. We will focus on the following aspects: (i) which intramolecular potentials are needed; (ii) are the internal degrees of freedom reproduced; (iii) is the "chirality" of the peptide reproduced; (iv) how do the two sets of atomistic sampling—with long range exclusions (set ex) and without (set noex)—compare; (v) do we reproduce potentially existing correlations of internal degrees of freedom; and (vi) does the resulting set of potentials yield the correct "overall" conformational sampling?

Intramolecular potentials. A total of six bond and seven angle potentials between CG beads are needed. Fig. 3 shows—using the distributions of angles at the first $C\alpha\beta$ bead as an example—that the Boltzmann inverted distributions can be successfully used as potentials in the CG simulations. Note that at this point only the CG^{noex} results are discussed (dashed line in Fig. 3). For the bonds and the other angles the agreement between atomistic and CG sampling is qualitatively very good (see Fig. 3a). As far as torsions are concerned, it
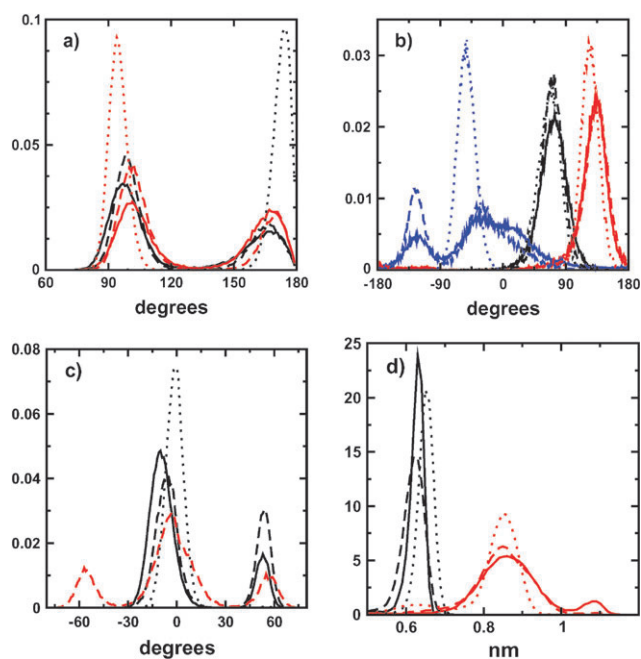

Fig. 3 Distribution for selected CG degrees of freedom from atomistic (full line), CG^{ex} (dotted line) and CG^{noex} sampling (dashed line). (a) Angles involving Phe_1: NH_3–$C\alpha\beta_1$–Phe_1 in black and Am–$C\alpha\beta_1$–Phe_1 in red. (b) Proper dihedral angles along the backbone $C\alpha\beta_1$–Am–$C\alpha\beta_2$–COO in black and NH_3–$C\alpha\beta_1$–Am–$C\alpha\beta_2$ in red, and involving a side chain Phe_1–$C\alpha\beta_1$–Am–$C\alpha\beta_2$ in blue. (c) Improper dihedral angle centered on $C\alpha\beta_1$; results omitting improper dihedral potentials in red. (d) Intra-peptide distances between the two termini in black and the two side chains in red.

needs to be tested which types and how many dihedral potentials are needed to reproduce backbone conformations of the atomistic sampling and to preserve the chirality of the backbone. This question affects the torsion around the $C\alpha\beta_1$–Am and the Am–$C\alpha\beta_2$ bonds, the question whether an improper dihedral is needed is relevant for the two $C\alpha\beta$ beads. It turns out that it is necessary to define two torsion potentials along each $C\alpha\beta$–Am bond, one involving the terminal bead and one involving the sidechain bead. Those are NH_3–$C\alpha\beta_1$–Am–$C\alpha\beta_2$ and Phe_1–$C\alpha\beta_1$-Am–$C\alpha\beta_2$ in the case of $C\alpha\beta_1$. Fig. 3b shows that the dihedral distributions are well reproduced with this approach (for the sake of clarity only three dihedrals are displayed, the one left out behaves qualitatively the same). In terms of which potentials are required to fully capture the peptide conformations the last aspect missing is the question of chirality. Fig. 3c shows that a special improper torsion potential (taking care of the out-of-plane motion of this bead with respect to its neighbors) is needed at the $C\alpha\beta$ beads. The black solid line corresponds to the distribution of this out-of-plane dihedral angle (defined as the angle between the $C\alpha\beta_1$–NH_3–Am and NH_3–Am–Phe_1 planes) sampled by the atomistic simulation. The distribution is clearly not symmetric with respect to zero which can be attributed to the chirality of the underlying atomistic peptide. On the contrary, the dashed red line, which shows the CG sampling setting the improper dihedral potentials to zero, is clearly symmetric, thus the CG peptide behaves achiral. The atomistic sampling can only be recovered using such an improper

torsion potential (which was obtained through Boltzmann inversion) in the CG simulation (dashed black line).

After having defined which types of interactions we need to properly describe the peptide, we compare two different approaches for the generation of bonded potentials. The sets were derived from the atomistic sampling with or without long range exclusions, and labeled as CG^{ex} and CG^{noex}, respectively (see section 2.2: Covalent/bonded interaction potentials, above). The ex approach refers to the case where atomistic conformational simulation to determine covalent interaction functions was performed with exclusions between the atoms corresponding to peptide beads further than three CG bonds apart, *i.e.* only those interactions are present that are to be replaced by bonds, angles or torsions on the CG level. First it was tested whether the atomistic simulation with long range exclusions leads to sampling of completely different conformations (compared to the normal atomistic sampling) by comparing a large number of conformational properties. It turned out that the conformations sampled are qualitatively very similar, however the weights with which various minima are sampled are in fact different between the ex and noex sampling (data not shown). So the question is, which sampling scheme results in CG Boltzmann inverted potentials that finally succeed better at reproducing the "real" peptide conformations (after nonbonded interactions between beads separated by more than three CG bonds are added, see section 2.3: Nonbonded interaction potentials, above). Fig. 3 clearly shows that the CG^{noex} potentials better reproduce the sampling of individual degrees of freedom.

In general, we observe that the distributions of all single degrees of freedom are reproduced very well in the sense that the peaks observed in the atomistic sampling are also found in the CG simulation. This means that the CG potentials do not by construction forbid sampling of relevant conformational minima. It can however be seen that not all peaks in the distributions are sampled with the same weight in the atomistic and the CG sampling.

Correlations of intramolecular degrees of freedom. An important step for the model validation is to check how far the assumption of uncorrelated intramolecular degrees of freedom is realistic, and whether the sets of potentials determined from atomistic sampling with or without exclusions will be capable of reproducing the peptide conformations on the coarse level. To analyze in more detail the performance of the two CG models, correlations between several intramolecular degrees of freedom were calculated and compared with the atomistic sampling. In the atomistic sampling, we observed correlations between angular and torsional degrees of freedom and between angles and torsions with the out-of-plane motion of the $C\alpha\beta$ beads denoted by the improper angle. The correlations between degrees of freedom can be visualized in two-dimensional graphs where the free energy (*i.e.* the Boltzmann inverted distribution) is displayed as function of the two (intramolecular) degrees of freedom. Fig. 4 shows the correlation between two intramolecular degrees of freedom centered around the first $C\alpha\beta$ bead, namely the angle NH$_3$–C$\alpha\beta_1$–Phe$_1$ and the improper dihedral angle (C$\alpha\beta_1$–NH$_3$–Am–Phe$_1$) for atomistic sampling (panel a), CG^{ex} sampling (panel b), CG^{noex} sampling
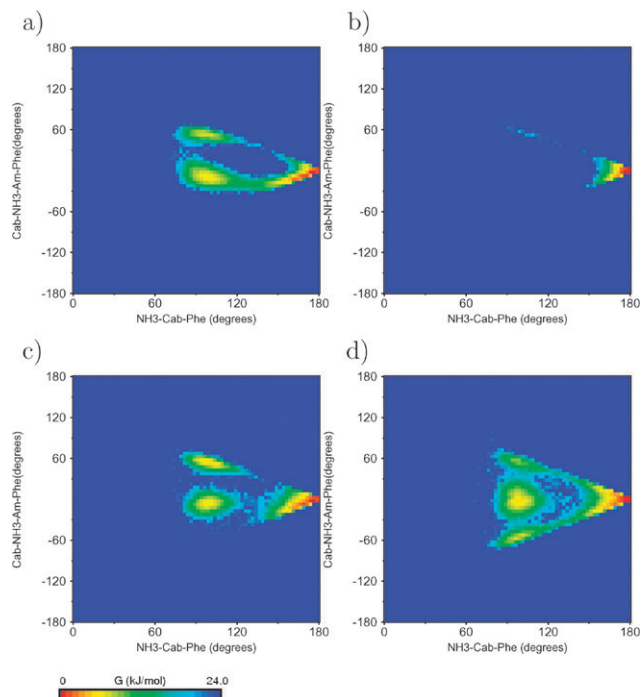

Fig. 4 Atomistic and CG sampling of the free energy landscape as a function of the CG degree of freedom: NH$_3$–C$\alpha\beta_1$–Phe$_1$ angle *versus* C$\alpha\beta_1$–NH$_3$–Am–Phe$_1$ improper dihedral angle. Results from different models: atomistic (a), CG^{ex} (b), CG^{noex} (c) and CG^{noex} omitting improper dihedral potentials (d). The free energy is given by $-k_BT[\ln P(i, j) - \ln P_{min}]$ where $P(i, j)$ denotes the probability of finding a conformation with values (i, j) for two select degrees of freedom and P_{min} is the population probability of the global minimum. A $\sin(\theta)$ normalization was applied to angles between two vectors. Angle values is degrees and free energy in kJ mol^{-1}.

(panel c), and CG^{noex} sampling when no improper dihedral potential was applied (panel d). (Note, that other pairs of intramolecular degrees of freedom, also those involving the second $C\alpha\beta$ bead behave qualitatively identically.) Comparison of panels (a) and (c) show clearly, that the full set of CG intramolecular potentials is capable of reproducing the correlations between degrees of freedom observed in the atomistic simulation. Fig. 4d again shows the importance of applying an improper dihedral potential to preserve the stereochemistry of the peptide. Also for this property, the CG^{noex} potentials appear to be more appropriate to describe the conformational equilibrium of the peptide system in water solutions than CG^{ex}.

A final test for the set of intramolecular CG potentials is the comparison between atomistic and CG sampling using an independent conformational property that has not been set *via* a corresponding potential. Fig. 3d shows the intra molecular distances between the termini and the phenyl groups, again obtained from atomistic sampling and from CG sampling with the CG^{ex} and the CG^{noex} sets of potentials. In general, both CG simulations reproduce the atomistic distributions qualitatively quite well. Again the CG^{noex} set of potentials yields a better agreement with the positions of the maxima of the atomistic distributions, in particular for the Phe$_1$–Phe$_2$ distance (red lines) (where the sampling with CG^{ex} misses the long distance tail of the distribution completely).

3.2 Peptide–peptide pair interaction and peptide self-assembly

In this section we validate the quality of the nonbonded interactions used for the CG dipeptide and investigate the effective peptide–peptide attractive interaction relevant for dimerization in dilute solution, as well as peptide self-assembly at a finite concentration. We first discuss the pair potential of mean force between two dipeptides in water obtained by constraint simulations in which the amide–amide distance was held fixed at distances ranging between 0.35 and 1.6 nm. The results of the CG model are compared to those of the corresponding all-atom model. In making this comparison, two questions are important. First, we ask if the range and depth of the attractive interaction observed with the detailed-atomistic model is reproduced by the coarse-grained model. Second, we ask if distance distributions between CG groups obtained with the CG model agree with the corresponding distributions obtained from all-atom simulations at the level of dimers as well as within larger aggregate structures.

Atomistic and CG pair potential of mean force. Fig. 5 shows the pair PMF as a function of the intermolecular amide–amide distance for two dipeptides in water obtained from all-atom explicit solvent simulations and simulation with the two CG models. In comparing the range of the attractive and repulsive parts of the PMF we observe good agreement between the atomistic and CG models. Also with respect to the depth of the attractive minimum, the two models are in reasonable agreement down to distances of about 0.8 nm. Below this distance, the all-atom model shows a maximum and a minimum, which are not reproduced with the CG model. The minimum corresponds to a direct hydrogen bonding interaction between the two peptide groups. To better quantify the discrepancy between the two PMF curves, we calculated the corresponding peptide–peptide association constant, $K_a = 4\pi \int_0^{R_{max}} r^2 \exp[-V_{PMF}(r)/k_B T] \, dr$ with $R_{max} = 1.6$ nm. The difference between the two models corresponds only in a small increase (6%) in the value of the peptide association constant in favor of CG model with respect to the atomistic values. The overall good agreement indicates that the approach used here to derive nonbonded bead–bead interaction potentials from all-atom simulations of the corresponding chemical moieties (see section 2.3: Nonbonded interaction potentials) does not
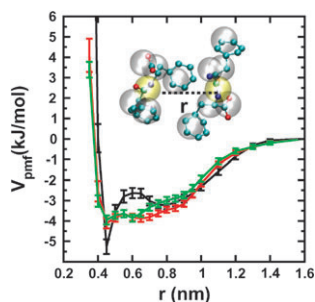
introduce unphysical behavior in the association free energy of the diphenylalanine pair.

Fig. 6a shows distance distributions between the side chains on the two peptides for selected amide–amide constraint distances. The all-atom distributions are qualitatively reproduced with the CG model; *i.e.* the locations of the peaks agree, whereas deviations are observed in the peak heights. Distribution functions of the mutual dipeptide orientations are shown in Fig. 6b for selected amide–amide constraint distances. In particular at close distance (0.45 nm), the peptides preferentially align antiparallel driven by the more favorable electrostatic interaction between the charged end groups. The CG model reproduces this behavior.

Note that the CG distributions, displayed Fig. 6, were obtained with the CG^{noex} intramolecular potentials, the CG^{ex} potentials showed qualitatively the same behavior with slightly worse agreement with the atomistic distributions (data not shown).

Based on the results presented in Fig. 5 and 6 we conclude that the CG model reproduces thermodynamic as well as structural features of pair association in good agreement with the detailed-atomistic model. This thermodynamic and structural agreement of peptide–peptide interaction confirms that the CG approach, which is based on small-molecule analogues of the CG beads, does not introduce unphysical conformations which dominate the behavior of the CG model.

Self-assembly of diphenylalanine in aqueous solution. To further examine the CG model, we performed MD simulations of self-assembly in a 0.6 M dipeptide solution. Because we want to compare the aggregated structures with predictions of
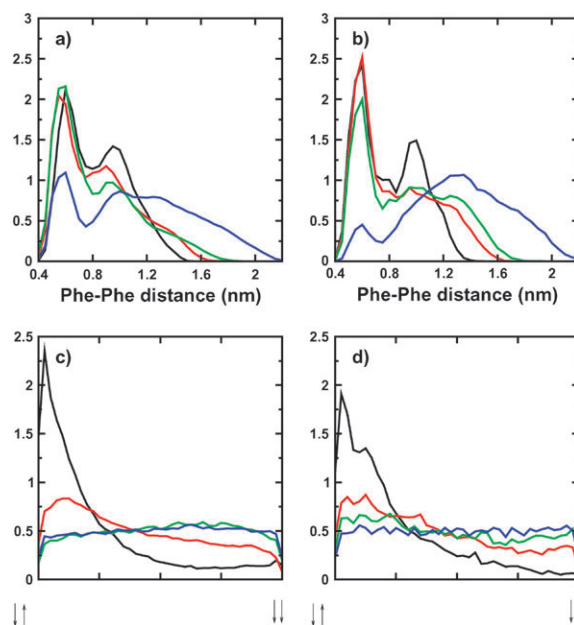

Fig. 6 Characteristic distribution for two peptides in water solution at a constraint distance of 0.45 nm (black), 0.65 nm (red), 0.80 nm (green) and 1.20 nm (blue). Top: intermolecular distance (nm) between phenyl rings from atomistic (a) and CG^{noex} sampling (b). Bottom: relative peptide orientations from atomistic (c) and CG^{noex} sampling (d). To define the relative peptide orientation, the scalar product between the unit vectors connecting the peptide termini was used.

Fig. 5 Potential of mean force between two peptides in water solution as a function of the distance between the central amide groups. Results from atomistic (black), CG^{ex} (green) and CG^{noex} (red) sampling are reported.

all-atom simulations, we chose a relatively small system size (96 dipeptides and 9000 water molecules). Comparison with experiments performed on this system[36–38] requires much larger system sizes and will be topic of a future publication. Both, CG and detailed-atomistic, systems were simulated starting with an initially random distribution of dipeptides upto the formation of an aggregate.

As a criterion for how fast the system approaches an equilibrium structure, we looked at the convergence of inter-molecular distance distributions as a function of time. In the atomistic simulation, 25×10^6 MD time steps are required to converge the distribution functions; 10^6 MD steps are required in the CG simulation. In the atomistic system, 25×10^6 integration time steps correspond to 50 ns. The time scale covered by the CG simulation is larger than this due to the use of softer interaction potentials and the implicit solvent representation. We have not estimated the time scale in the CG system which could be done, for example, by a time mapping of the CG peptide diffusion constant on the atomistic system. To describe the observed aggregate, we select radial distribution functions (RDFs) of sidechain and termini distances. Note that only results from CG^{noex} are described further. Fig. 7 shows RDFs for selected pairs in the self-assembled peptide solution modeled with the all-atom and CG simulations. The insert shows snapshots of the self-assembled aggregates obtained with the atomistic and CG models. The pair distribution of side chain distances (left panel in Fig. 7) observed with the atomistic model shows two peaks whose locations are reproduced with the CG model. The CG model, however, shows stronger density fluctuations over the full range of distances, indicating that the CG model forms peptide clusters that are overall more densely packed. The RDF of the charged end groups (right panel in Fig. 7) shows a slightly more complicated structure. The atomistic RDF corresponding to COO–NH$_3$ pairs shows three peaks below 0.6 nm, followed by a minimum and a broad maximum around 1.3 nm. The CG model does not reproduce the first two peaks below 0.6 nm because these are features related to differences in COO–NH$_3$ hydrogen bonding arrangements that the CG model cannot resolve. At distances larger than 0.6 nm the CG model reproduces the features of the all-atom RDF while, in the overall picture, the COO–NH$_3$ correlations are again too strong due to the higher density of packing. The COO–COO RDF has a peak at 0.55 nm in, both, the all-atom and CG system. At larger range, the CG model shows stronger correlations.

Re-insertion of atomistic details. To evaluate the accuracy of our mapping scheme, the atomistic details were reintroduced into the equilibrated CG peptide system. To verify that the relaxation of the atomistic system did not lead away from the underlying coarse-grained conformations the average displacement was calculated. An average root mean square (RMS) deviation of 0.06 nm was calculated for the seven CG beads between the CG and the corresponding equilibrated atomistic configuration. We re-introduce atomistic details also on the 2-peptide and 96-peptide system. As an example, the insert in Fig. 5 shows for the re-inserted atomistic details in the 2-peptide system. After 2 ps MD equilibration simulations without any position restraints of the 96-peptide system in water solution, the RMS deviation of all CG beads between the start CG configuration and the corresponding atomistic back-mapped configuration is 0.22 nm. The so-obtained atomistic configuration (after resolvating with explicit water) was used as a starting structure for a short atomistic simulation to compare the stability of the CG and atomistic aggregate. We observe that the sidechain and termini RDFs start to converge to the atomistic curve (see full line in Fig. 7) after 5 ns.

The success of the back-mapping procedure and the comparable structure and stability of the CG and atomistic clusters illustrates that the CG model developed here can be used together with the atomistic model in multiscale modeling procedures where the two models are hierarchically linked. CG simulations provide a computationally inexpensive route to modeling self-assembly at mesoscopic length and time scales, while the inverse-mapping procedure, combined with a short all-atom MD simulation, permits to return back to atomistic scales when required. To give an idea of the dynamic speed-up achieved with the CG model, we compare the CPU time to observe aggregate formation with the two models based on the 96-peptide system. Aggregate formation at the CG level is at least 1000 times faster than at the atomistic level. Because the CG interactions are softer than the atomistic ones, this performance can be further improved by increasing the CG time step.
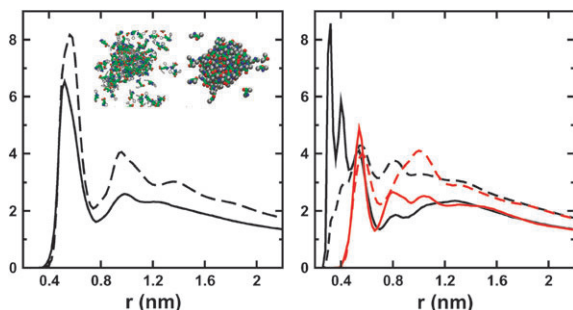
4. Methods/computational details

All simulations were performed using the GROMACS suite of programs (version 4.0).[39–42]

4.1 Atomistic simulations

All the simulations were performed in a periodic cubic box with a length longer than twice the cut-off. A cut-off was used for the Lennard-Jones interactions of 1.4 nm and a long range dispersion correction was applied for energy and pressure. The particle mesh Ewald method[43] was employed to treat Coulomb interactions, using a switching distance of 1.0 nm, a grid of 0.12 nm and a beta value of 3.1 nm^{-1}. Constant pressure p and temperature T were maintained by coupling the system to an external bath at 1 bar and 300 K, using Parrinello–Rahman

Fig. 7 Radial distribution functions of the solution of 96 peptides: atomistic (full line) and CG^{noex} (dashed line) results. Left: inter-molecular side-chain distance. Insert shows a atomistic (left) and CG^{noex} (right) snapshot of the system colored according to Fig. 1. Right: intermolecular termini distance: COO–NH$_3$ curve in black; COO–COO curve in red.

barostat[44] and Nosé–Hoover thermostat,[45,46] respectively. The pressure coupling time was 3 ps and the isothermal compressibility 4.5×10^{-5} bar^{-1}. The bond distances and the bond angle of the solvent water were constrained using the SETTLE algorithm.[47] All other bond distances were constrained using the LINCS algorithm.[48,49] A leap-frog integrator with a integration time step of 2 fs was used. The GROMOS 53a6 force field[50] was used for the peptide and SPC/E[51] model for water. For the backmapping procedure, a force constant of 20 000 kJ mol^{-1} nm^{-2} was used for virtual site position restraints and the steepest descent algorithm for the energy minimization.

4.2 Coarse-grained simulations

Tabulated potentials were used for bonded and nonbonded interactions between the beads. The simulations were performed at constant volume and at 300 K. As volume, the average volume of the corresponding atomistic simulation was used. A leap-frog stochastic integrator with a integration time step of 0.002 ps was used.[52] The inverse friction constant was set to 1.0 ps.

4.3 Boltzmann inversion of bond, angle and dihedral distributions

For the tabulated interaction potentials and forces one can in principles use the numerically Boltzmann inverted distributions [according to eqn (1–3)] and the corresponding derivatives, which are however usually rather noisy (in particular the forces). For the bonded potentials we therefore numerically fitted the Boltzmann inverted distributions with a polynomial (up to order 8 if necessary) and thus obtained smooth tables for both potentials and forces.

4.4 Calculation of potentials of mean force

The potential of mean force, V_{PMF}, between two molecules was calculated from n distance constraint simulations, using the following equation:

$$V_{\mathrm{PMF}}(r) = \int_{r_{\mathrm{m}}}^{r} \left[\langle f_{\mathrm{c}} \rangle_s + \frac{2k_{\mathrm{B}}T}{s} \right] \mathrm{d}s + C, \qquad (5)$$

where k_{B} is the Boltzmann constant and T is the temperature. f_{c} is the constraint force between the two solute centers of mass and r_{m} the maximum distance between the two mass centers. When the distance between the two mass centers is constrained, free rotation of the solute–solute connecting vector remains possible and larger volume elements are sampled at larger distances. This leads to an entropic contribution to the average constraint force that must be subtracted out. The second term of eqn (5) takes care of this contribution. To keep the distance between the two molecules fixed a linear constraint solver (LINCS)[48,49] was used.

For the nonbonded potential procedure, initial configurations were generated by moving the two solutes apart with a speed of 0.001 nm ps^{-1} *in vacuo* starting from an equilibrated conformation at a distance of 0.28 nm. Configurations each 0.02 nm were solvated and simulated. A total of 60 constraint simulations were performed. Each simulation was performed for 10 ns at the atomistic level.

Because the effective potentials between two ions are very close to a Coulomb interaction after 1.0 nm, we assumed beyond 1.2 nm:

$$V_{\mathrm{PMF}}(r) = \frac{q_1 q_2}{4\pi\varepsilon_0\varepsilon_{\mathrm{r}} r}, \ r > 1.2 \, \mathrm{nm} \qquad (6)$$

where q_1 and q_2 are the charges of the ions, ε_0 the dielectric permittivity of vacuum. For the relative dielectric permittivity, ε_{r}, a value of 72 (value for the SPC/E water model) was used. Thus, the constant C in eqn (5) is given by the Coulomb potential at 1.2 nm for the ion-ion interaction, while it is zero in all the other cases.

For the peptide–peptide pair interactions, the potential of mean force was calculated from 19 distance constraint simulations according to eqn (5). The distance between the center of mass of the central amide group (corresponding to the Am bead) was used as coordinate. The constraint distances varied from 0.35 to 1.6 nm. Atomistic simulations at short (<0.45 nm) and large (>1.2 nm) distance were performed for 20 ns, while 100–200 ns simulations were performed for each distance between 0.45 and 1.2 nm. The corresponding CG simulations were performed for 20 000 ps.

5. Summary and conclusions

Multiscale simulations of self-assembly and the formation of hierarchical structures of chemistry-specific peptide systems are currently infeasible because simple—yet sufficiently specific—CG models are unavailable. Therefore, an urgent need exists to examine coarse graining methodologies for these types of systems in order to enable large-scale simulations capable of probing mesoscopic time and length scales. In the context of peptide self-assembly, important features that CG peptide models should realistically describe are conformational flexibility and nonbonded interactions, which relate to aspects of structure and thermodynamics, respectively. Finding an optimum balance in describing these two aspects with CG models still remains a significant challenge.

As a first step in this direction, this paper presents a methodology to derive a CG model for diphenylalanine in aqueous solution based on detailed-atomistic molecular dynamics simulations. Hydrophobic dipeptides are the smallest biological entities forming self-assembled hierarchical structures in aqueous solution. This paper describes the development of a flexible CG model with the aim to reproduce the conformational sampling and thermodynamic association properties of the corresponding atomistic model. No assumptions are made on the functional form of the bonded and nonbonded interaction potentials which are based on potential of mean force calculations. The effects of aqueous solvation on the conformational sampling and on the nonbonded interactions are effectively included in these potentials.

The bonded part of the CG force field has contributions of bonds, angles and torsion angles, which are assumed to be uncoupled. We discuss the validity of this assumption and show that with the chosen CG mapping scheme, the statistical correlations between the CG degrees of freedom, as observed in detailed-atomistic simulations, can nevertheless be accurately reproduced. With the resulting set of bonded potentials,

and a 4-body potential (improper dihedral) that fixes the peptide's chirality, the conformational sampling of the CG model is in excellent agreement with the conformational sampling of the corresponding atomistic model. To describe the CG nonbonded interactions between the CG peptide beads, we calculate pair potentials of mean force from detailed-atomistic simulations using molecule pairs representative of the CG beads. This approach, combined with the assumption of pair-wise additivity of the resulting implicit-solvent nonbonded bead–bead interactions yields a complete description of the effective nonbonded interaction between CG peptides. This building block approach to describe the non-bonded peptide–peptide interaction has been validated by comparing the thermodynamic association constant of a di-peptide pair obtained with the atomistic and CG models. The observed agreement is within 6%.

The dipeptide model resulting from our CG parameteriza-tion scheme has been used to simulate the self-assembly of dipeptides at finite concentration is solution. With the CG model, the computational efficiency increases by at least three orders of magnitude in comparison to the detailed-atomistic force field model. The success of the CG model relies not only on the ability to correctly sample intra and the intermolecular conformations, but also on the possibility to re-insert the atomistic details. Using an inverse-mapping procedure, we generate representative atomistic coordinates from the CG trajectory with a maximum root mean square deviation of 2 Å (for self-assembled aggregates based on 96 peptides). The successful forward and backward mapping approaches makes the model suitable to be used in a multiscale strategy, where scale-hopping between the mesoscopic and the atomistic length scales is a crucial aspect.

Finally, we point out that with the coarse graining approach described in this paper uncertainties related to the accuracy of the atomistic force field model are automatically carried over to the CG model. Hence, the ability of the CG model to uncover mechanistic details of self-assembly and structure formation in the experimental dipeptide solution hinges upon this question. Given, however, that the CG model is poten-tially capable of sampling the mesoscopic domain probed in experiments, this question can be turned around; *i.e.* CG force field models may be used to validate atomistic force field models, which, when used by themselves only, cannot be validated in particular cases, simply due to sampling problems.

Acknowledgements

References

1 S. J. Marrink, D. P. Tieleman and A. E. Mark, *J. Phys. Chem. B*, 2000, **104**, 12165–12173.
2 C. F. Lopez, S. O. Nielsen, P. B. Moore and M. L. Klein, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 4431–4434.
3 M. Muller, K. Katsov and M. Schick, *Phys. Rep.*, 2006, **434**, 113–176.
4 G. S. Ayton, W. G. Noid and G. A. Voth, *Curr. Opin. Struct. Biol.*, 2007, **17**, 192–198.
5 P. L. Freddolino, A. Arkhipov, A. Y. Shih, Y. Yin, Z. Chen, and K. Schulten, in *Coarse-Graining of Condensed Phase and Biomolecular Systems*, ed. G. A. Voth, Chapman and Hall/CRC Press, Taylor and Francis Group, 2008.
6 B. J. Reynwar, G. Illya, V. A. Harmandaris, M. M. Müller, K. Kremer and M. Deserno, *Nature*, 2007, **447**, 461–464.
7 N. F. A. van der Vegt, C. Peter and K. Kremer, in *Coarse-Graining of Condensed Phase and Biomolecular Systems*, ed. G. A. Voth, Chapman and Hall/CRC Press, Taylor and Francis Group, 2008.
8 M. Praprotnik, L. Delle Site and K. Kremer, *Annu. Rev. Phys. Chem.*, 2008, **59**, 545–571.
9 S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman and A. H. de Vries, *J. Phys. Chem. B*, 2007, **111**, 7812–7824.
10 L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman and S. J. Marrink, *J. Chem. Theory Comput.*, 2008, **4**, 819–834.
11 J. Michel, M. Orsi and J. W. Essex, *J. Phys. Chem. B*, 2008, **112**, 657–660.
12 P. J. Bond and M. S. P. Sansom, *J. Am. Chem. Soc.*, 2006, **128**, 2697–2704.
13 H. D. Nguyen and C. K. Hall, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 16180–16185.
14 P. Derreumaux and N. Mousseau, *J. Chem. Phys.*, 2007, **126**, 025101–10717.
15 G. Bellesia and J. E. Shea, *J. Chem. Phys.*, 2007, **126**, 245104–10717.
16 S. Brown, N. J. Fawzi and T. Head-Gordon, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 10712–10717.
17 J. D. Honeycutt and D. Thirumalai, *Proc. Natl. Acad. Sci. U. S. A.*, 1990, **87**, 3526–3529.
18 S. Izvekov and G. A. Voth, *J. Phys. Chem. B*, 2005, **109**, 2469–2473.
19 J. Zhou, I. F. Thorpe, S. Izvekov and G. A. Voth, *Biophys. J.*, 2007, **92**, 4289–4303.
20 C. Peter, L. Delle Site and K. Kremer, *Soft Matter*, 2008, **4**, 859–869.
21 A. P. Lyubartsev and A. Laaksonen, *Phys. Rev. E*, 1995, **52**, 3730–3737.
22 F. Muller-Plathe, *ChemPhysChem*, 2002, **3**, 754–769.
23 I. Bahar and R. L. Jernigan, *J. Mol. Biol.*, 1997, **266**, 195–214.
24 V. Tozzini, W. Rocchia and J. A. McCammon, *J. Chem. Theory Comput.*, 2006, **2**, 667–673.
25 P. Das, S. Matysiak and C. Clementi, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 10141–10146.
26 A. Liwo, P. Arlukowicz, C. Czaplewski, S. Oldziej, J. Pillardy and H. A. Scheraga, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 1937–1942.
27 V. A. Harmandaris, N. P. Adhikari, N. F. A. van der Vegt and K. Kremer, *Macromolecules*, 2006, **39**, 6708–6719.
28 W. Tschop, K. Kremer, J. Batoulis, T. Burger and O. Hahn, *Acta Polym.*, 1998, **49**, 61–74.
29 C. F. Abrams and K. Kremer, *Macromolecules*, 2003, **36**, 260–267.
30 Q. Sun and R. Faller, *Macromolecules*, 2006, **39**, 812–820.
31 B. Hess, C. Holm and N. F. A. van der Vegt, *J. Chem. Phys.*, 2006, **124**, 164509–61.
32 A. P. Lyubartsev, *Eur. Biophys. J.*, 2005, **35**, 53–61.
33 N. Basdevant, D. Borgis and T. Ha-Duong, *J. Phys. Chem. B*, 2007, **111**, 9390–9399.
34 B. Hess, S. Leon, N. F. A. van der Vegt and K. Kremer, *Soft Matter*, 2006, **2**, 409–414.
35 W. Tschop, K. Kremer, O. Hahn, J. Batoulis and T. Burger, *Acta Polym.*, 1998, **49**, 75–79.
36 E. Gazit, *FASEB J.*, 2002, **16**, 77–83.
37 C. H. Görbitz, *Chem.–Eur. J.*, 2001, **7**, 5153–5159.
38 M. Reches and E. Gazit, *Science*, 2003, **300**, 625–627.
39 H. J. C. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Comm.*, 1995, **91**, 43–56.

40 E. Lindahl, B. Hess and D. van der Spoel, *J. Mol. Model.*, 2001, **7**, 306–317.
41 D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark and H. J. C. Berendsen, *J. Comput. Chem.*, 2005, **26**, 1701–1718.
42 B. Hess, C. Kutzner, D. van der Spoel and E. Lindahl, *J. Chem. Theory Comput.*, 2008, **4**, 435–447.
43 T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.
44 M. Parrinello and A. Rahman, *J. Appl. Phys.*, 1981, **52**, 7182–7190.
45 W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695–1697.
46 S. Nosé, *Mol. Phys.*, 1984, **52**, 255–268.
47 S. Miyamoto and P. A. Kollman, *J. Comput. Chem.*, 1992, **13**, 952–962.
48 B. Hess, H. Bekker, H. J. C. Berendsen and J. G. E. M. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.
49 B. Hess, *J. Chem. Theory Comput.*, 2008, **4**, 116.
50 C. Oostenbrink, A. Villa, A. E. Mark and W. F. van Gunsteren, *J. Comput. Chem.*, 2004, **25**, 1656–1676.
51 H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *J. Phys. Chem.*, 1987, **91**, 6269–6271.
52 W. F. van Gunsteren and H. J. C. Berendsen, *Mol. Simul.*, 1988, **1**, 173–185.