JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Coarse-Grained Simulations of Protein Backbone Dynamics. 1. Local Sterics Define the Dihedral Angles

Andreas Wagenmann* and Tihamér Geyer*

Zentrum für Bioinformatik, Universität des Saarlandes, D-66041 Saarbrücken, Germany

ABSTRACT: Here, we present a coarse-grained model targeted for implicit solvent simulations of unfolded or intrinsically disordered proteins. The hierarchical model with its nonspherical building blocks allows one to reproduce the local dynamics of the backbone with simple harmonic bonds and steric collisions between a small number of atoms at the correct off-center positions on the building blocks. Here in part 1, we also describe the implementation of the global shape of the protein chain and the extended local interactions that add a first secondary structure bias, which will subsequently be augmented by additional hydrophobic interactions, hydrogen bonds, and dipole dipole couplings along the backbone. Due to its hierarchical setup, the model has a near-atomistic resolution on the local scale and the overall numerical efficiency of a coarse-grained model such that even long protein chains can be simulated efficiently.

INTRODUCTION

Proteins, the "machines" of a cell, have a quite interesting structure. On the one hand, they are linear chains of amino acids that are assembled by the ribosomes according to a template taken from the linear DNA code, while, on the other hand, these essentially one-dimensional chains have to fold into a specific three-dimensional structure to become functional. They vary in sizes from less than 100 amino acids to protein complexes containing more than 1 million of these building blocks, and they perform numerous tasks such as enzymatic and catalytic activities, transport of ions and molecules into and out of a cell, energy conversion, immuno response, and structural functions. All these diverse structures and functions are realized from a rather small set of 20 different amino acids which only differ in their side chains but have the same backbone structure. These observations naturally trigger the reductionist's question of how one can understand, for example, the folding of an arbitrary protein chain into its respective 3D structure from this limited set of building blocks.

Another vital aspect of protein folding is when folding goes wrong. Misfolding and subsequent aggregation for example are hallmarks of a whole class of prominent diseases like Alzheimer's and Parkinson's diseases or spongiform encephalopathies.[1,2] These pathologies share a common structural scheme in which peptides or proteins fold into β-rich structures and then agglomerate into ordered aggregates. The Alzheimer A_β peptides, for example, are in a mainly α-helical state as part of the amyloid precursor proteins (APP). Excised from the APP and floating in the extracellular fluid, they can adopt a β-strand configuration and form ordered oligomers.

While experimental methods can provide accurate data about native folds and stable intermediates, they can give only relatively crude and indirect information about the pathways and dynamics of the actual folding process. Therefore, computational models are the method of choice to investigate the dynamics of protein folding and fill in the gaps between the experimentally accessible (meta-)stable snapshots. All-atom force fields provide the most detailed representations, and currently, they allow one to routinely simulate events in the 10 μs range for small systems on typical computer clusters and up to some 100 ms on specialized hardware.[3,4] However, α-helices and β-hairpins form in 0.1 to 10 μs, and small proteins fold within tens of microseconds.[5] This means that the folding dynamics of only small proteins can be investigated. The long time scales that are required to simulate the folding of larger proteins or multiprotein aggregation can, however, be reached with coarse-grained (CG) models which trade resolution for speed. In CG models, the proteins are described with a reduced number of interaction sites. These reduced models have already shown some success in reproducing correct structures for small proteins or aggregation events on large scales.[6−11] This demonstrates that a limited set of interaction sites is sufficient to reproduce overall properties of protein folding and protein−protein association.

Many current coarse-grained modeling schemes for protein dynamics build on the idea of "super atoms" which represent a number of real atoms or chemical groups. Effective interactions are then calculated between the centers of these still spherical super atoms.[9,12] However, using such effective spheres to describe an (often nonsymmetric) group of atoms limits the modeling flexibility and is the main obstacle for making these methods even more "coarse" by grouping more and more atoms. Also, nontrivial multibead potentials may be required to capture correlations, e.g., between dihedral angles of adjacent residues.

A different approach is presented here. Instead of uniting atoms until a sufficient degree of simplification and efficiency is achieved, we start from a detailed description of the local interactions between atoms close by and add more and more simplified hierarchies of longer ranged interactions. Additionally, we do not use enlarged spherical "atoms" but allow our building blocks to have arbitrary shapes. This ansatz builds on the observation that at the lowest level, steric clashes between a few protein backbone atoms confine the allowed regions of the

Received: June 29, 2012
Published: September 18, 2012

ACS Publications © 2012 American Chemical Society

dx.doi.org/10.1021/ct3005529 | J. Chem. Theory Comput. 2012, 8, 4732−4745

Ramachandran dihedral space. Ho et al. for example found that for a generic residue the Ramachandran space is delimited by clashes between three atoms only.[13,14] Any further layers of interactions added onto this basic scaffold then only serve to stabilize folded structures and to reproduce the correct residue-specific secondary structure propensities. These additional layers include steric, hydrophobic, and electrostatic dipole interactions between neighboring residues on the first level; helix-breaking clashes between residues one turn apart as another weaker bias for secondary structure formation; or globally visible shapes to prevent the self-overlap of the peptide chain. This model suffices to simulate folding and dynamics of individual secondary structure elements like helices and β-turns. For the assembly of these structural elements, another layer will eventually be added that describes the side chains in more detail. In this hierarchic approach, each additional layer can be added and parametrized independently on top of the already existing model. The result is a model which has a near atomistic resolution on the local scale and the global numerical efficiency of a coarse-grained model.

For simulations with our model, an implicit-solvent Langevin propagation scheme[15] is used so that, with realistic diffusion coefficients for the individual building blocks, the dynamics are reproduced on the correct time scales. The first version of the model as presented here focuses on the residue-specific formation of secondary structure elements and less on the dynamics of protein folding. Therefore, hydrodynamic interactions, which induce velocity correlations between the individual building blocks but do not change the energetics, are not included yet. It is clear, though, that they may affect the actual folding dynamics as was recently discussed by Frembgen-Kesner and Elcock.[16,17] With our "Brownmove" simulation package,[18] however, it is straightforward to also include them in the next iteration when the side chains will be implemented in more detail, too.

The important new aspects of our approach are thus the use of arbitrarily shaped building blocks that allow for a great flexibility in the modeling, the hierarchic interactions, which keep both the conceptual and the computational complexity at a manageable level, and the use of an implicit solvent propagation scheme for realistic time scales for folding and agglomeration dynamics.

Our hierarchic coarse-grained protein backbone model is described in two parts. The first part, given in this publication, focuses on the basic concept and on the steric structure of the backbone. In this part also, the implementation within our "Brownmove" simulation package, the propagation scheme, and the reference data for the parametrization are explained. The second part will then show how the next layers consisting of electrostatic and hydrophobic interactions and hydrogen bonds both allow and bias the folding of stable secondary structure elements.

METHODS

Comparison to other Coarse-Grained Approaches. Coarse-grained, i.e., simplified representations of proteins and polymers on various levels of accuracy have been used for quite some time now. Mostly, these models were tailored for a specific project or question, and it was only recently that the first systematic, general-purpose models were developed. Some prominent examples of such general-purpose models are the united-residue (UNRES) model developed in the Scheraga group,[19] the multiscale coarse-graining approach of Voth and Izvekov,[20] the MARTINI force-field for coarse-grained molecular dynamics simulations with explicit water,[12] and, most recently, a "generic coarse-grained model for protein folding and aggregation" by Bereau and Deserno.[21] Though the actual implementations of these models are quite different from each other, the common basic idea was to develop a coarse-grained model starting from the atomistically detailed representation that would reproduce the behavior of the protein chain with a reduced number of interaction sites. This selection of the interaction sites is the most crucial step because it determines the form of the interaction potentials and thus the efficiency and accuracy of the model. As can be seen in the following, there is a certain level of complexity required to faithfully model the behavior of the protein backbone. This means that a simple geometric representation as in the UNRES case goes with more complex multibead potentials, while the more detailed building blocks of our model allow for simple binary interactions.

In the UNRES model, the development of which was started nearly two decades ago,[9] one interaction site is placed on the peptide bond halfway between adjacent C_α's and a second one on the side chain. For an efficient numerical representation, transformed coordinates are used on the basis of the vectors between neighboring C_α's and between the C_α's and their respective side chains. The interactions have various functional forms ranging from simple binary interactions over dipole−dipole terms up to multiparticle dihedral angles. A part of the interaction potentials was implemented as statistical potentials that were determined from structures from the protein data bank (PDB),[22] while others are restricted free energy profiles obtained by integrating out atomistic degrees of freedom in the detailed representation.[9,19] Due to this smoothing of the energy landscape, the dynamics takes place faster than with the respective "rough" atomistic model. In the propagation, an additional diffusive contribution of random kicks and a friction term is added which is scaled with the solvent accessible surface of each bead. Thus, a bead inside a folded protein only sees the forces from the surrounding residues, whereas a (partly) solvent-exposed bead is additionally kicked around by the implicit solvent. This model allows one to run unbiased protein folding simulations.[7,23]

To mimic this coarse-graining related speedup in fully atomistic molecular dynamics simulations without actually coarse-graining, Wei and Wang recently investigated the idea to smoothen the energy landscape by damping the short ranged interactions.[24] However, here the problem also arises that the speedup is not easily quantified, and therefore rate or time constants can at best be estimated.

Also aimed at protein folding is the recent force field of Bereau and Deserno[21] for coarse-grained implicit solvent molecular dynamics simulations. In their model, the backbone is represented with three beads at quasi-atomistic resolution, while also only one bead is used for the side chain. In order to describe secondary structure formation, the bead sizes were optimized such that the Ramachandran angle distributions are reproduced best. Secondary structures are stabilized by hydrogen bonds and simplified dipole−dipole terms. Further residue specificity was added via the Miyazawa−Jernigan hydrophobicity scale.[25]

Whereas the two just mentioned force fields focus on the protein backbone, a general "multiscale coarse-graining" scheme to convert an atomistic representation consistently onto a simplified model was developed by Voth and co-

workers.[26] Once a mapping is defined that relates the groups of atoms to their respective coarse-grained substitutes, the masses of these and the interaction potentials can be calculated from the atomistic force field. Depending on the mapping operator, arbitrarily complex many-particle potentials may arise. This procedure has been applied to a wide variety of problems ranging from simple liquids[20] to protein stability.[27] In approaches like this one, where "less important" degrees of freedom are traced out, not only the choice of the coarse-graining "chunks" may affect the results but also which atomistic force field is selected as a starting point.[28]

A mathematically more simple strategy was used for the MARTINI force field.[12] Here, three to four close-by atoms are grouped together into one superatom. These superatoms, which are then classified as polar, nonpolar, apolar, or charged, interact via the usual energy terms of classical molecular dynamics, i.e., bond and angle terms and nonbonded Lennard-Jones and Coulomb interactions. In the MARTINI model also the solvent is coarse-grained, and therefore the internal and the diffusional dynamics are treated in a consistent manner. However, all dynamics are too fast, too, due to the smoothed energy landscape. This force field has been used very successfully for lipid−water systems or membrane permeation studies, whereas for stable protein secondary structures additional stabilizing constraints have to be added which then prevent, e.g., unbiased *ab initio* protein folding.

From a historical perspective, the oldest CG models for protein folding were one-bead-per-residue models with potentials that drive the protein chains into the folded conformation. With these so-called Gō models, many important principles of protein folding could be understood as, e.g., the concept of an energy funnel or the two major pathways of either hierarchical folding or hydrophobic collapse.[29] An interesting variant of the interaction potential was introduced recently by Wolff et al., which is based on the connectivity of the individual residues.[30] Going beyond the limitations of these simplistic models, more detailed representations were set up that then additionally could describe dynamic processes. Most simple models consisted of two beads, where one bead, the "C_α" bead, represented the backbone and the other bead, the side chain. However, only with four to six spherical beads is there enough level of detail on the backbone to provide realistic conformational sampling in the Ramachandran space. Four-bead models used the beads {N,H}, {C_α,H_α}, {C,O}, and {R} for the side chain,[21,31−37] while six-bead models were built from {N}, {H}, {C_α}, {C}, {O}, and {R}.[38] In some models, H and O were used only for hydrogen bond formation.[39,40] Favrin et al. used a second side chain bead for proline (C_δ) instead of the amide hydrogen and a φ-dihedral fixed at $-65°$ to model its helix-breaking properties.[40] In most models, the interaction site R was centered on the first side chain atom C_β. A different approach was taken by Takada et al.[32,33] and Smith and Hall,[34] who used a side-chain-dependent position. For a more detailed overview, see, e.g., the recent review by Tozzini.[41]

This brief overview illustrates that there were two complementary trends in coarse-grained modeling. On the one hand, there are the "real" coarse-graining approaches where detail is removed by grouping more and more atoms, whereas the historically older "fine-graining" ansatz was to start from the most simple representation and to add more detail once the current representation was not sufficient any more. As explained in the following, our model combines ideas from both approaches. From the coarse-graining approach, we take the concept to group atoms into rigid building blocks, whereas the hierarchic addition of interactions to initially simple "beads" in an implicit solvent resembles the "fine-graining" approach.

The main difference, however, to any of the previous approaches is that in our model the rigid subunits need not be simple spherical or ellipsoidal particles but can have any shape. These shapes can even be different for each of the considered interactions. At first, this may sound unnecessarily complex, but as we will show, this flexibility allows one to use simple and intuitively understandable interaction potentials, and the various types of interactions can be implemented and parametrized independently. With this, our model has a similar scope and resolution as the molecular dynamics based united atom approaches. However, in contrast to these, the propagation of the building blocks is based on Brownian motion, i.e., the microscopic picture of the macroscopic process of diffusion. This has the consequence that time scales are represented correctly in our model when the correct diffusion coefficients are provided, and the smoothed energy landscape does not lead to accelerated dynamics. This can be understood when one considers that on the local, microscopic scales, the random thermal motion that mimicks the heat-bath of the surrounding molecules dominates the trajectories. This constant back-and-forth has, together with the Stokesian friction term, a similar effect as the microscopic roughness of the energy landscape in atomistic models. Consequently, in contrast to the UNRES force field, we do not scale the random and friction terms with the solvent accessible surface but include them at the same strength no matter whether the beads are fully solvated or completely buried within a protein or complex. Our motivation for this is based on the interpretation of the random kicks and the friction forces of the implicit solvent as a thermostat. Naturally, this neighborhood thermostat does not end at the surface of a protein, and also the inner parts of a protein have their local fluctuating thermostat. Now when coarse-graining is performed, the beads inside a protein then have a different environment than in an atomistic representation. The artificial spatial discretization into rigid building blocks means that the energy dissipating quasi-continuous deformations as they can occur in the atomistic model are replaced by frictionless displacements along the boundaries of the rigid subunits. The random kicks and the friction against the background then qualitatively recover the thermostat effect within the protein bulk. Effectively, the individual coarse-grained building blocks behave as if they would move on an energy landscape with a fluctuating roughness on an atomistic scale, and thus the time scales are not accelerated as we could show for a first handmade model of a short peptide.[15] For this peptide, we could obtain roughly the same power spectrum of the vector from the first to the last bead from our coarse-grained Langevin dynamics simulations as from an all-atom molecular dynamics simulation with explicit water.

On this issue there is, however, still some room for debate, as recent atomistic molecular dynamics simulations of short peptides showed.[42] In these simulations, the mass of the solvent molecules was varied in order to separate the influence of the external friction, i.e., between solvent and protein, from the protein-internal friction. The results demonstrated that both contributions affect the dynamic behavior of folding proteins.

Recent simulations also confirmed that hydrodynamic interactions should even be included inside a multibead protein

model when the relative timing of the internal and the external dynamics as well as the speed ratio between translational and rotational diffusion is important.[17,18]

Modeling Protein Structure and Interactions. The coarse-grained protein backbone model presented here was implemented within our "Brownmove" simulation package.[18] Thus, when we describe the implementation of our ansatz we will use the "Brownmove" terminology. However, any other simulation package could have been used, too, provided that it is flexible enough to model arbitrarily shaped objects and eccentrically attached bonds.

In "Brownmove", a flexible protein is implemented from rigid building blocks. Each of these "beads" is initially a shapeless container into which various types of interaction sites can be placed at arbitrary positions. This hierarchy is reflected in Brownmove in an object oriented fashion, where a "Protein" object contains one or multiple "Gestalt" objects (the "beads"), which in turn contain "Shape" objects holding the actual interaction sites (see Figure 1A). For any type of interaction,

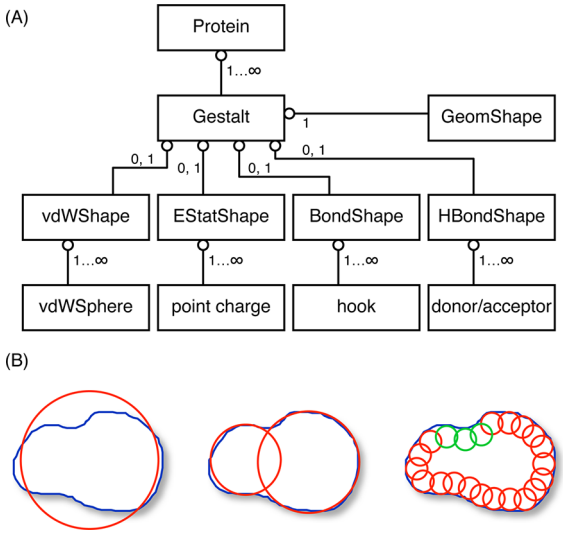

Figure 1. (A) Class structure of a protein as implemented in the "Brownmove" simulation package[18] with the interaction types that are relevant for the coarse-grained protein model presented here. A flexible protein contains multiple rigid "Gestalt" objects (the beads), which in turn contain objects for the required interaction sites. For the short ranged van der Waals interactions these are, e.g., spheres, while electrostatic interactions are implemented with point charges. The "GeomShape" object implements the propagation algorithms and is therefore always present. This hierarchic approach allows one to model a protein at different resolutions as sketched in part B. This can range from a single van-der-Waals sphere up to a large number of small spheres with different interaction parameters to model, e.g., a hydrophobic patch as indicated by the red and green spheres.

there is one "Shape" class; i.e., brownmove has, for example, an "EStatShape" class which handles electrostatic interactions, a "vdWShape" class for effective short ranged interactions, or the "BondShape" which allows one to connect two "Gestalt" beads via elastic bonds. Currently not used for this protein model are the shapes for hydrodynamic interactions or external fields.

For the propagation of such coarse-grained structures, both the conventional Brownian dynamics as well as a recently introduced Langevin propagation scheme[15] can be used (see below). Beyond the set of features required for the protein model, Brownmove can also handle many-particle scenarios with constant or variable particle numbers,[43] treat hydrodynamic interactions efficiently,[44] and be easily extended with other types of interactions or reactions.

The protein model and the necessary tools for setup and data extraction from the trajectories are freely available for academic use together with the underlying Brownmove package.[18] These include scripts that convert PDB files into protein definition files for Brownmove and also tools to, e.g., extract coarse grained snapshots or the atom positions from the trajectory files. Such reconstructed atomistic trajectories can then be analyzed with standard tools for molecular dynamics simulations.

Spatial Shapes: van-der-Waals Interactions. In our model, the spatial shape of a bead is described by one or more "van-der-Waals" spheres with an effective potential for the short ranged interactions. The spheres belonging to one "vdWShape" object may overlap and have arbitrary radii. Different "colors" for each sphere allow to have different interaction potentials for different pairs of spheres to describe, e.g., hydrophilic or hydrophobic surface parts. In the simplest approximation, a protein is modeled with a single sphere, which would be implemented in brownmove as "one vdWSphere in the vdWShape of the single Gestalt of the Protein object". A more sophisticated description would use a small number of "vdWSpheres" of different sizes, while a semirealistic spatial shape can be implemented with many small spheres placed along the surface of the protein (see Figure 1B). It should be noted that different resolutions can be used within the same simulation. For the effective short ranged interactions between the spheres, a 6−12-Lennard-Jones type potential is used.

$$E_{\mathrm{vdW}}(d_{ik}) = C_{12}\left(\frac{r_0}{d_{ik} + dr}\right)^{12} + C_6\left(\frac{r_0}{d_{ik} + dr}\right)^{6} \quad (1)$$

The distance d_{ik} is measured between the respective surfaces of sphere i on one Gestalt and sphere k on the other Gestalt, i.e., $d_{ik} = r_{ik} - (R_i + R_k)$ with the distance r_{ik} between the centers of the spheres which have the radii R_i and R_k. The parameter r_0 describes the range of the potential independently of the sphere radii and dr allows one to shift the potential. This can be used, e.g., to have an interaction energy of $E_{\mathrm{vdW}} = 1\ k_{\mathrm{B}}T$ for $d_{ik} = 0$, i.e., when the spheres touch. For numerical stability, the potentials can be linearized below a certain minimal distance.

Bonds and Bond-Like Interactions. The individual beads can be connected with elastic springs hooked up at arbitrary positions within the beads. The standard bond is described by a harmonic potential with a spring constant κ and a rest length L.

$$E_{\mathrm{HO}}(r_{ik}) = \frac{\kappa}{2}(r_{ik} - L)^2 \quad (2)$$

In our protein model, such harmonic springs are used for the chemical bonds along the backbone and also for preserving angles, as in brownmove, no three-body interactions are used.

To efficiently model the local interactions between specific pairs of adjacent residues, special bond types with van-der-Waals and Coulomb potentials were implemented. When a bond of these types is used between two beads, the interactions between the corresponding shapes are skipped. Thus, local interactions between charges or van-der-Waals spheres are implemented with these bonds, whereas for the non-neighbor-specific nonlocal interactions, point charges and spheres are used. By setting the respective interaction strengths to zero,

dx.doi.org/10.1021/ct3005529 | J. Chem. Theory Comput. 2012, 8, 4732−4745

these special bonds can also be used to switch off interactions between selected pairs of beads.

Implicit Solvent Propagation Scheme. When the individual building blocks of a protein model are large enough compared to the water molecules, an implicit solvent propagation scheme can be used. The most famous one is Brownian dynamics (BD), which has become a workhorse technique with the Ermak-McCammon algorithm.[45] BD is based on Einstein's insight that the individual collisions between the large particle of interest and the many small solvent molecules can be replaced by a velocity dependent friction term plus random kicks with a vanishing average and a temperature dependent strength.[46] With this ansatz, the many-particle Newton equation for, e.g., a protein in an explicit water box collapses to a one-particle Langevin equation for the protein alone.

$$\frac{dv}{dt} = \frac{1}{m}(F + f_r - \gamma v) \tag{3}$$

Here, v and m are the velocity and mass of the protein, F denotes the sum of all external forces onto the protein, f_r is the random kicks, and γ is the friction coefficient of the protein. Assuming that F and f_r are constant over a short time interval Δt, this equation can be integrated analytically to give the velocity $v(\Delta t)$ and the displacement $\Delta x(\Delta t)$ after one time step Δt when $v(0) = v_0$ initially and $F_t = F + f_r$:

$$v(\Delta t) = \frac{F_t}{\gamma} + \left(v_0 - \frac{F_t}{\gamma}\right)\exp\left[-\frac{\gamma \Delta t}{m}\right] \tag{4}$$

$$\Delta x(\Delta t) = \frac{F_t}{\gamma}\Delta t - \frac{m}{\gamma}\left(v_0 - \frac{F_t}{\gamma}\right)\left(1 - \exp\left[-\frac{\gamma \Delta t}{m}\right]\right) \tag{5}$$

For rotation, an analogous set of equations is used. The standard BD propagation is obtained from these equations in the limit of either a large time step or a small mass such that the velocity relaxation time $\tau = m/\gamma \ll \Delta t$. Then, v is directly proportional to F_t, and the displacement increases linearly with Δt.

$$\Delta x(\Delta t) = \frac{F_t \Delta t}{\gamma} \tag{6}$$

As can be seen, BD involves the two approximations that (i) the solvent molecules are much smaller than our particles of interest, and (ii) that the propagation time steps are much longer than the velocity relaxation time of the smallest particle. The first condition stems from the implicit solvent, whereas the second approximation is for numerical simplicity only. As explained in detail previously,[15] the Langevin dynamics algorithm of eqs 4 and 5 is as efficient as BD but can also be used on the small scales of residue-level simulations of peptides and proteins.

Reference Structures. Structural reference data were used extensively throughout the parametrization of our force field. For this, structures were retrieved from the PDB database[22] with the criterion that the resolution was ≤1.8 Å, the free R value was ≤0.25, and the B value was ≤30. PDB files in which two chains had the same chain identifier, multiple entries existed for any backbone atom, or iCodes were used, which may cause ambiguities regarding relative residue positions, were removed. This resulted in a total of 5370 structures. Backbone hydrogens were added using the Open Babel[47] software package, and the structures were normalized to the default PDB naming scheme. The above-mentioned criteria were then applied residue-wise, and three sets of valid residues were extracted. (i) The *backbone set* consists of all residues in which all backbone atoms satisfy the quality criteria. This set was used for information about dihedral distributions. (ii) For the *extended backbone set*, those residues were removed from the *backbone set* in which—for all nonglycine residues—the first side chain atom did not satisfy the criteria. For proline, all heavy side chain atoms had to be valid. (iii) Consensus coordinates for the heavy atoms of the side chains were extracted from the *side chain set*. It consists of those residues in which all backbone and all heavy side chain atoms satisfied the criteria.

From these sets, secondary structure elements were identified using the STRIDE[48] software package. Extracted types were α and 3_{10} helices and β-turns and strands. We also merged β-strands that were hydrogen bonded by at least two hydrogen bonds to get β-sheets and hairpins. In hairpins, the two connected strands belong to the same chain. STRIDE does not always include flanking residues, i.e., the outer residues that provide the first/last enclosing hydrogen bond, into α-helices. This depends on how close the respective dihedrals match the helical conformation. For 3_{10} helices, STRIDE just gives the inner residues. In both cases, we added the flanking residues to the helix. For β-turns, STRIDE does include the flanking residues. To investigate how a secondary structure element affects neighboring residues, we also identified up to two neighbors on each side of a secondary structure element if these were valid according to the quality criteria given above. Note that for these residues, their direct neighbors further away from the secondary structure element had to be classified as "coil" according to STRIDE; otherwise they were excluded. This was done to prevent another adjacent secondary structure element affecting the conformation of the respective residue.

RESULTS AND DISCUSSIONS

This section explains how our coarse-grained model of the protein backbone is built up. The first step of the hierarchic setup is the definition of the building blocks (the "beads") and how these are connected. The next layer then consists of clashes between individual atoms of adjacent beads. These clashes limit the conformational flexibility of the underlying bare scaffold to the allowed regions in the Ramachandran dihedral space. A third layer of globally visible shapes prevents the self-overlap of the protein chain, while some further "extended local" atom−atom collisions between more distant residues add a first bias for secondary structure formation. The model, however, will only be complete when the nonsteric interactions like electrostatics, hydrophobicity, and hydrogen bonds are also added. These will be explained in the second part of the publication. Consequently, the "really interesting" applications cannot be shown yet, and the main results of this part explain the hierarchical modeling concept and demonstrate that the local dynamics are reproduced correctly, i.e., that we now have a scaffold onto which the longer-ranged, residue-specific interactions that determine for example folding can be attached to.

Building Blocks and Bonds of the Backbone. The protein backbone consists of repeated units of N−C_α−C atoms. The peptide bond which connects these repeated units is planar and rather rigid due to its partial double bonds. The complete flexibility of the backbone thus comes from the rotations around the bonds next to the C_α atoms. This is sketched in Figure 2A, where the red arrows indicate the rotatable bonds of

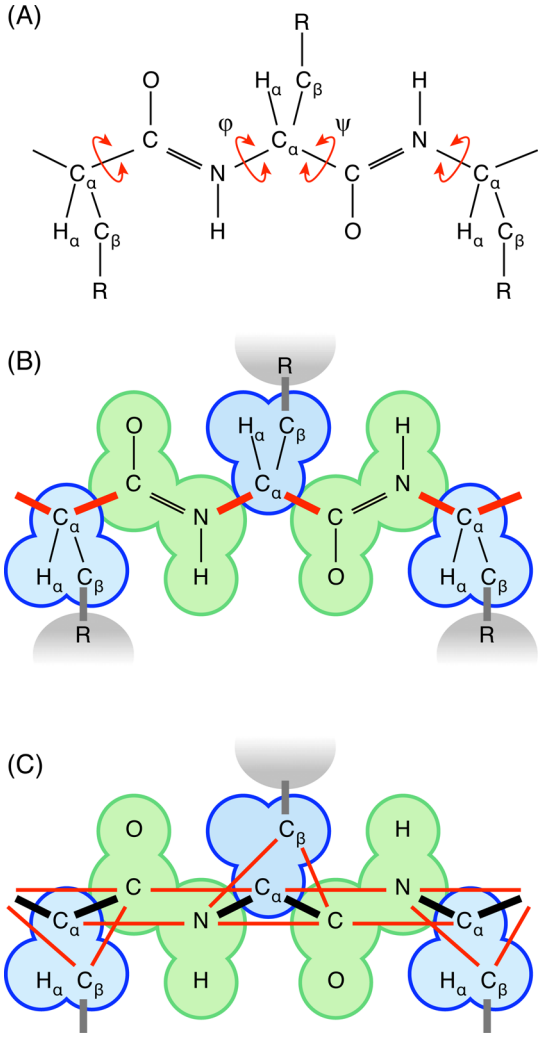

Figure 2. Local interactions along the backbone. The backbone as sketched in panel A is partitioned at the rotatable bonds, which are indicated by the red arrows. This leads to the two rigid backbone building blocks of the peptide bond (green) and the C_α–C_β part (light blue) as shown in panel B. These rigid subunits are connected by off-center bonds (thick red lines). The respective angles between these bonds and the two shapes that they connect are preserved by additional bonds as indicated by the thin red lines in panel C. Note that at this level the two building blocks do not have any spatial shape yet; the blue and green shapes are only drawn to better visualize the partitioning of the backbone.

the C_α. We therefore take the peptide group as one rigid building block and the C_α plus its hydrogen and some part of the side chain as a second rigid subunit. Additionally, the N- and C-terminal groups are implemented as building blocks, too. The bonds that connect these building blocks are then attached at the respective (off-center) positions of the C and the N atoms of the peptide group and at the C_α, respectively (see Figure 2B). This backbone structure is identical for all residues. They only differ in their side chains and, potentially, additional residue-specific constraining bonds. For proline, e.g., the φ dihedral is restricted by the bent-back side chain. This is implemented in our model by an additional bond hooked up at the position of the C_δ side chain atom and the N of the preceding peptide group. For the most simple case of glycine, there is no side chain and in our model the respective C_α group consists of the C_α and the two attached hydrogens.

As in Brownmove only two-body interactions are implemented, the respective bond angles are fixed with additional bonds as indicated in Figure 2C. It is easy to see that the resulting structure built from the rigid subunits, the chemical bonds, and the angle-preserving bonds reproduces the conformational flexibility of the original protein backbone with its (yet unconfined) rotations around the Ramachandran angles.

The bond lengths and the respective angles were taken from Engh and Huber,[49] who analyzed a large data set of high-resolution structures. On the basis of these values, the relative positions of the atoms within the C_α unit, the peptide bond, and the N- and C-terminal building blocks were determined. Each of these building blocks then translates and rotates as a rigid body with no internal dynamics under the influence of the total force and torque from the various individual force contributions. The individual atom positions are only required to specify at which point relative to the origin of the building block a certain interaction acts. The internal coordinate system of the C_α units is determined by e_1 as the normalized vector from the C_α to the following C atom. For e_2, the orthonormalized projection of the vector from the C_α to the preceding N is used, and the third unit vector is $e_3 = e_1 \times e_2$. The origin of this internal coordinate system of the C_α unit is placed at the position of the C_α . The resulting atom positions within this reference frame are listed in Table 1. These relative coordinates can then later also be used to recover an atomistic representation from the coarse-grained simulation snapshots. When for example a bond is attached at the C_β atom, then in the model the bond is hooked up at the respective internal coordinates. The internal coordinate system of the peptide

Table 1. Positions of the Required Atom Positions in the Rigid Building Block Units of Our Protein Model[a]

building block	atom	coordinates [nm]	remark
C_α	C_α	0.0/0.0/0.0	
	C_β	−0.0526/−0.0779/−0.121	not for GLY, PRO
	H_α	−0.05040/−0.03445/0.08909	-"-
	$H_{\alpha 1/2}$	−0.05034/−0.03446/∓0.08911	for GLY
	C_β	−0.0526/−0.0631/−0.1206	for PRO (averaged)
	C_γ	−0.1296/0.0393/−0.1985	-"-
	C_δ	−0.1306/0.1650/−0.1200	-"-
	$C_{\gamma 1}$	−0.0014/−0.2217/−0.1200	for VAL, TRP, PHE, TYR
	$C_{\gamma 2}$	−0.2047/−0.0783/−0.1231	-"-
	$C_{\gamma 1}$	−0.2058/−0.0832/−0.1191	for ILE
	$C_{\gamma 2}$	0.0044/−0.2195/−0.1245	-"-
	$O_{\gamma 1}$	−0.0104/−0.0151/−0.2404	for THR
	$C_{\gamma 2}$	−0.2013/−0.0885/−0.1215	-"-
peptide bond	C	−0.02374/−0.03687/0.0	
	O	−0.08897/0.06752/0.0	
	N	0.1104/−0.03687/0.0	
	H	0.1634/−0.1217/0.0	
C-term	C	−0.042/0.0/0.0	
	$O_{x1/2}$	0.01575/∓0.1086/0.0	
N-term	N	0.0/0.0/0.0	H atoms are ignored

[a]These atom positions are used to hook up the various interactions within the internal coordinate systems of the rigid building blocks. The definitions of the respective internal coordinate systems are given in the text.

dx.doi.org/10.1021/ct3005529 | J. Chem. Theory Comput. 2012, 8, 4732–4745

bond unit is centered on the center of mass of the C, O, N, and H atoms and defined from the normalized vector from the C to the N atom, the orthonormal component of the vector from the C to the O, and the cross product of these two. For the C terminus, the vectors from the C to the midpoint between the two O_x's and the vector from O_{x1} to O_{x2} are used, centered on the center of mass. For the N terminus, the hydrogens are ignored and the N defines the origin.

The bonds that connect adjacent building blocks are listed in Table 2. Basically, there are two types. The first connects the N-terminus or a peptide unit to the following C_α block, whereas the other links the C_α unit to the subsequent peptide unit or the C-terminal. A third bond type connects the bent-back side chain of a proline residue to the N atom of the preceding peptide group. The other distances given in Table 2 all lie within one of the rigid blocks and were only used to determine the relative positions of the atoms within their respective units.

Table 2. Bonds of Our Protein Model[a]

atom pair	bond length [nm]	remark
$N-C_\alpha$	0.145	
$C_\alpha-C$	0.152	
$N-C_\delta$	0.147	for PRO
$C_\alpha-C_\beta$	0.153	distance within building block
C−O	0.123	-″-
C−N	0.134	-″-
N−H	0.1	-″-
$C_\alpha-H_\alpha$	0.108	-″-

[a]The bonds are hooked up in the building blocks at the respective positions of the two connected atoms (see Table 1). The two atoms are given in the sequence in that they occur along the protein chain starting from the N-terminus. This table lists both the actually implemented bonds and those that were only used to determine the relative positions of the atoms in the building blocks.

Table 3. Angle Defining Bonds of Our Protein Model[a]

defining atoms	angle	bond length [nm]	remark
$N-C_\alpha-C$	111.2°	0.245	
$C_\alpha-C-O$	120.8°	0.239	
$C_\alpha-C-N$	116.2°	0.244	
$C-N-C_\alpha$	121.7°	0.244	
$N-C_\alpha-C_\beta$	110.4°	0.244	
$C_\beta-C_\alpha-C$	110.5°	0.250	
$C_\alpha-C-O_{x1/2}$	118.0°	0.236	for C-term
$H-N-C_\alpha$	119.1°	0.212	
$N-C_\alpha-H_\alpha$	108.6°	0.206	
$H_\alpha-C_\alpha-C$	108.6°	0.212	not GLY, PRO
$N-C_\alpha-H_{\alpha1/2}$	108.6°	0.206	for GLY
$H_{\alpha1/2}-C_\alpha-C$	108.6°	0.212	-″-
$C-N-C_\delta$	126.8°	0.250	for PRO
$N - C_\delta - C_\gamma$	101.5°	0.224	-″-
H_1-N-C_α	111.2°	0.203	for reconstruction of the N-term
$H_{2/3}-N-C_\alpha$	108.6°	0.200	-″-
$O_{x1}-C-O_{x2}$	124.0°	0.217	for reconstruction of the C-term

[a]The atoms are given in the sequence in that they occur along the protein chain starting from the N-terminus. The bond is attached to the two outer of the three atoms that define an angle. The last three angles are not represented by bonds but are only required to reconstruct the atomistic representation from the coarse-grained model.

In addition to these "chemical" bonds between subsequent units, another nearly 20 additional bonds are required to define the various bond angles which are not "buried" within the rigid building blocks. These angle-defining bonds with their lengths and hook-up positions are listed in Table 3.

The spring constants for all bonds were determined by running simulations of 10-residue polyalanines with the bonds as the only interactions. From these simulations, where a rather short time step of 6 fs was used, the variances of the bond lengths and angles were determined with respect to the spring constants. On the basis of these results, the spring constant for the bonds along the backbone was set to 4×10^4 kJ mol^{-1} nm^{-2}, for which the bond lengths fluctuated within 0.04 nm. For the angle constraining bonds, the same spring constant of 4×10^4 kJ mol^{-1} nm^{-2} kept the angles to within ±5°. For essentially all simulations reported in this publication, we left the time step at 6 fs, although tests indicated that longer timesteps can be used, too.

Local Steric Clashes. The above-defined model with its still shapeless building blocks and the various bonds can perform arbitrary rotations around the φ and ψ dihedral angles, whereas in a real protein the flexibility of the backbone is restricted by clashes between the nonpointlike atoms. Ramachandran et al.[50] were the first to identify the allowed regions in the dihedral space and to relate them to the observed secondary structures. For this they treated the atoms as hard spheres. Subsequent analyses that used slightly elastic van-der-Waals spheres for the atoms only showed minor changes in the locations of the clash-induced boundaries of the allowed regions. Recently, Ho et al. carefully analyzed a high resolution data set from the protein data bank and determined probability maps in the Ramachandran space for different types of residues.[13] By this they identified a simplified set of atom−atom clashes that can describe the allowed and the forbidden regions in the dihedral angle space. Furthermore, they showed how electrostatic dipole−dipole interactions further shape the observed angle distributions within the allowed regions.

For a generic residue, i.e., a residue with a side chain that consists of at least the C_β, the atoms that need to be considered are shown in Figure 3. Panel A shows a Ramachandran plot with the regions that are blocked and the respective pairs of colliding atoms. These clashes are shown in panel B in the context of the protein chain. Interestingly, the rather complex landscape of allowed and blocked regions of the dihedral angle space can be explained by collisions of only three atoms: the O and N of the peptide bond and the C_β. The C_α atom, which is the starting point of most other coarse-grained models, is not involved here. In our model, we implemented these steric clashes identified by Ho et al. as bonds with a short ranged van-der-Waals repulsion hooked up at the respective positions on the rigid building blocks where the real atoms would sit in a fully atomistic model. For efficiency reasons, bonds are used instead of globally visible van-der-Waals spheres. The softer $C_{\beta i}-O_i$ clash is treated as the other clashes.

The analysis by Ho et al. also showed that for the observed dihedral angle distributions only four different types of residues are required. In addition to the generic residue where the side chain consists of at least the C_β atom, the side-chain-less glycine, the proline, and the residue before a proline (the "preproline") are required. In the glycine case, the second H_α atom takes the place of the C_β of a generic residue, while in the

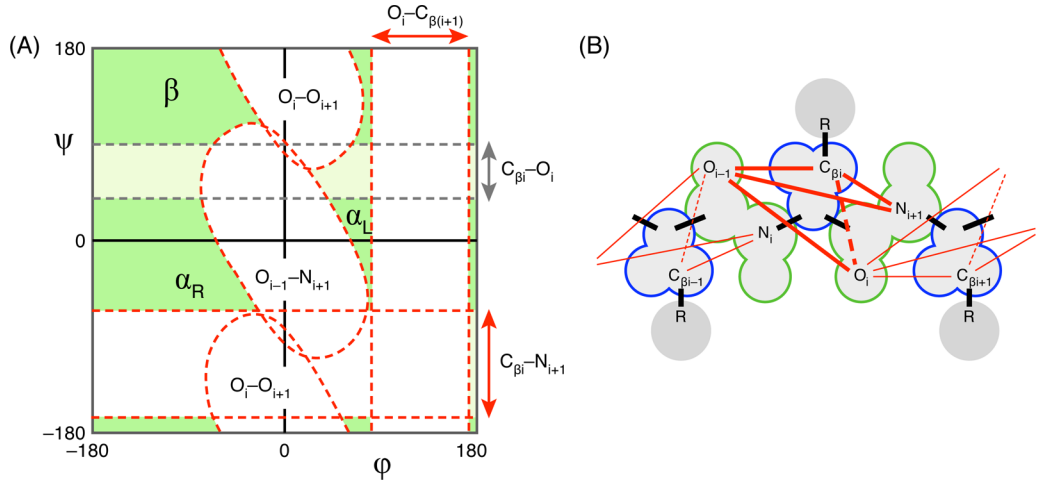

Figure 3. (A) According to the analysis by Ho et al.,[13] only a small number of local steric clashes between nearby atoms need to be considered to explain the observed dihedral distributions of the Ramachandran angles. For a generic residue, these are the pairs O_i–O_{i+1}, O_i–$C_{\beta(i+1)}$, $O_{(i-1)}$–$N_{(i+1)}$, and $C_{\beta i}$–$N_{(i+1)}$, where i denotes the index of the residue. The respective configurations that are blocked by these clashes are indicated by the broken red lines and the white areas. A softer clash which does not completely block the respective φ–ψ region occurs between $C_{\beta i}$ and O_i (gray lines, light green area). This figure is redrawn according to the work of Ho et al.[13] These local interactions are implemented in our model as eccentrically attached, repulsive van-der-Waals bonds as shown in panel B for C_α–C_β unit i and the adjacent peptide bonds which partly belong to residues $(i-1)$ and $(i+1)$, respectively (cf. Figure 2). The thin red lines indicate the clashes with the next C_α units to the left and to the right, respectively. Note that the outlines of the building blocks are shown only for illustration; they do not denote actually implemented spatial shapes.

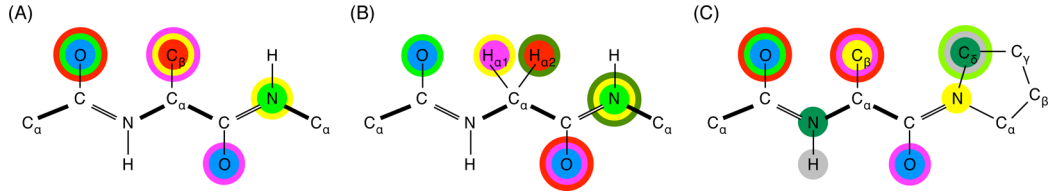

Figure 4. Overview over the implemented local interactions for (A) a generic residue, (B) glycine, and (C) preproline. Pairs of interacting atoms are indicated with a circle of the same color. This image is an (incomplete) graphical representation of Table 4. The rotatable bonds are indicated by thick lines.

proline case an additional restriction comes from the C_δ of the rigid side chain. These cases are sketched in Figure 4 while the complete list of local steric clash bonds of our model is given in Table 4.

As a starting point for the bond lengths of these local steric clashes between the backbone atoms, we used the sums of the respective atom radii from Word et al.[51] With simulations of GAG and GGG dipeptides, these values were optimized to best reproduce the sampling of the Ramachandran angles for the generic and the glycine residue classes. A second round of distance optimizations was performed with 12-residue poly-alanines for the generic class, polyglycines for the glycine case, and AAPAPAPAPAPAA polypeptides with independent analyses for the prolines and preprolines (here: alanines). Suitable interaction strengths were found to be C_{12} = 0.6 kJ mol^{-1} both for the hard and for the soft local clashes. The r^{-6} term was not used, i.e., C_6 = 0. For numerical stability, the potential was linearized for $E_{vdw} \geq 20$ kJ mol^{-1} (see eq 1).

From the simulations, we finally obtained the dihedral angle distributions of the four residue classes shown in Figure 5. In this figure, the black lines denote the limits identified by Ho et al.[3,14] The good agreement between the distributions from the simulations and the limits identified from high-resolution structures shows that our conceptually simple backbone model in which off-center bonds connect the building blocks such that the flexibility of the backbone comes from the rotations around the Ramachandran angles plus a small number of steric clashes between nearby atoms already suffices to map out the

Table 4. Local Atom–Atom Collisions Considered in Our Model[a]

atom pair	R_0 [nm]	residue types
O_i–$O_{(i+1)}$	0.360	any
O_i–$O_{x(i+1)}$	0.360	$(i+1)$ = C term
O_i–$C_{\beta(i+1)}$	0.355	$(i+1) \neq$ GLY, PRO
$O_{(i-1)}$–$N_{(i+1)}$	0.355	$(i+1) \neq$ PRO
$O_{(i-1)}$–$C_{\delta(i+1)}$	0.345	as previous, but for $(i+1)$ = PRO
$C_{\beta i}$–O_i	0.345	$i \neq$ GLY
$C_{\beta i}$–O_{xi}	0.345	$i \neq$ GLY to C term
$H_{\alpha1/2i}$–O_i	0.297	as previous, but for i = GLY, two bonds
$H_{\alpha1/2i}$–O_{xi}	0.297	i = GLY to C term, two bonds
$C_{\beta i}$–$N_{(i+1)}$	0.360	$i \neq$ GLY
$H_{\alpha1/2i}$–$N_{(i+1)}$	0.302	as previous, but for i = GLY, two bonds
$N_{(i-1)}$–$C_{\delta i}$	0.350	i = PRO
$H_{(i-1)}$–$C_{\delta i}$	0.285	-"-

[a]The residue with index i consists of the ith C_α atom, its side chain, the previous N, and the next C. For all these van-der-Waals bonds, C_{12} = 0.6 kJ mol^{-1} and dr = 0 were used (see eq 1). For a graphical illustration, see Figure 4.

dx.doi.org/10.1021/ct3005529 | J. Chem. Theory Comput. 2012, 8, 4732–4745

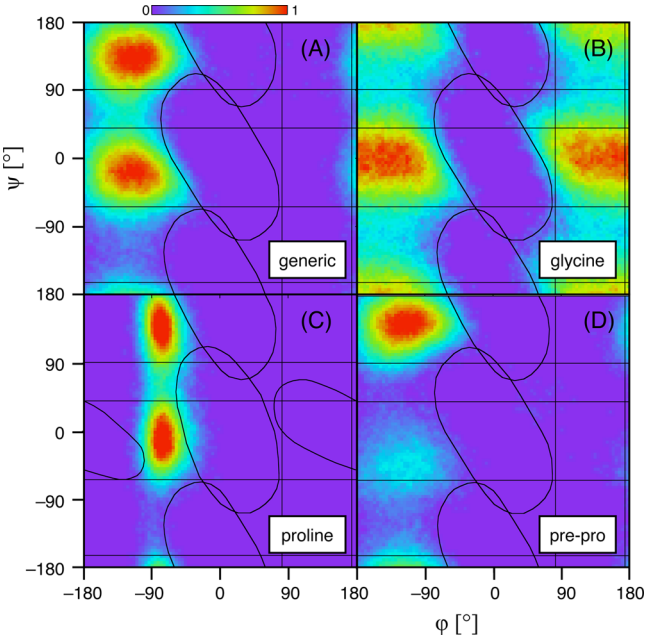

Figure 5. Distribution of dihedral angles for the different residue types from simulations of polypeptides with our coarse-grained model when only the bonds and the local steric clashes according to Table 4 are implemented. A square root scaling is used for the frequencies so that the low-occupancy regions can be seen better. All distributions are rescaled to fit into the range [0...1]. The black lines indicate the excluded regions due to local clashes between pairs of individual atoms as identified by Ho et al.[13,14] (also see Figure 3 and Table 4).

respective residue-specific allowed regions in the Ramachandran space. With this agreement on the single-residue level, extended secondary structures may already form, but there is no stabilization yet as all interactions (except for the bonds) are purely repulsive. Also, though our model with the bonds and the local sterics reproduces the allowed regions, the actual distributions that were extracted from, e.g., X-ray structures, show more details within the allowed $\varphi-\psi$ regions. In these data sets, the α region, for example, has a more diagonal shape due to electrostatic dipole interactions between adjacent residues. Consequently, at the current point the basic structure of our protein model is set up and defined, and all additional layers of interaction will then serve to stabilize and to bias secondary structure formation.

Global Sterics. The interactions defined so far are local, i.e., between atoms of adjacent building blocks. Thus, these local interactions do not affect the global folding and compaction of the protein when secondary and tertiary structures develop. For this, global interactions were implemented. As their purpose is to prevent a self-overlap between residues distant in sequence while the local dynamics are already well-defined, these globally visible shapes can be implemented rather coarse-grained with a small number of van-der-Waals spheres per residue. Using the smallest number of such globally visible interaction centers is important for performance reasons, because the effort to evaluate all forces between them scales quadratically with their total number. In detail, the peptide bond is modeled with two spheres which are placed at the van-der-Waals-radius weighted centers of the C−O and of the N−H atom pairs, respectively. With these two spheres, the peptide bond has a dumbbell-like spatial shape. For the C_α building block, one sphere was placed at the position of the C_α, while another sphere at the C_β position models the first part of the side chain. For glycine, the C_β sphere was omitted, and the C_α sphere was placed at the van-der-Waals-radius weighted center of the C_α and its two hydrogens, whereas for proline a larger sphere stands for the C_β, C_γ, and C_δ atoms. The resulting positions of these van-der-Waals spheres are listed in Table 5. Each of these spheres has a different van-der-Waals "color" index so that for each combination of spheres optimized distances can be used.

Table 5. Relative Positions of the Globally Visible van-der-Waals Spheres within the Internal Coordinate Systems of Their Respective Building Blocks[a]

vdW sphere	coordinates [nm]	remark
CO	−0.05684/0.0161/0.0	
NH	0.1296/−0.06771/0.0	
C_α	0.0/0.0/0.0	not for GLY
GLY-C_α	−0.02881/−0.01971/0.0	C_α, $H_{\alpha 1/2}$ for GLY
P-C_β	−0.0524/0.0780/−0.1209	avg. for polar residues
PRO-C_β	−0.1043/0.0471/−0.1464	planar configuration of PRO
ALA-C_β	−0.0527/−0.0772/−0.1206	
CYS-C_β	−0.0518/−0.0772/−0.1213	
PHE-C_β	−0.0524/0.0774/−0.1213	
ILE-C_β	−0.0537/−0.0789/−0.1215	
LEU-C_β	−0.0522/−0.0772/−0.1214	
MET-C_β	−0.0524/−0.0777/−0.1210	
TRP-C_β	−0.0523/−0.0776/−0.1212	
TYR-C_β	−0.0522/−0.0775/−0.1213	
VAL-C_β	−0.0533/−0.0793/−0.1214	

[a]For the polar residues ARG, LYS, ASN, GLU, GLN, HIS, SER, THR, and ASP, the same averaged position of the C_β sphere is chosen, whereas for the hydrophobic side chains, residue-specific positions are used.

These simplified globally visible spatial shapes should not interfere with the highly detailed interactions that define the local backbone dynamics. This is achieved with a feature of the Brownmove package that the globally visible van-der-Waals spheres are ignored when one or more van-der-Waals bonds exist between a pair of Gestalt objects. Consequently, the global spatial shapes are only relevant for residues beyond the range of the local steric interactions, i.e., beyond the dipeptide level.

Initial estimates for the interaction widths between the global shapes were obtained from the pairwise distance distributions extracted from the high-resolution reference data. The bond lengths were then adjusted such that the protein chain could still be folded into an α-helical structure, i.e., that the global steric shapes just start to overlap and only a little attraction by the hydrogen bonds would be needed to stabilize the helix. These optimized values of the interaction widths between the various pairs of global spheres are listed in Table 6. At the current interaction hierarchy level of the model, no hydrophilic or hydrophobic interactions are included. Correspondingly, only the repulsive r^{-12} term of the Lennard-Jones potential (see eq 1) is used with C_{12} = 0.6 kJ mol^{-1}. Again, dr = 0, and the potentials are linearized beyond 20 kJ mol^{-1}.

One way to demonstrate the effect of the global steric shapes is via the radial distance distribution between the C_α atoms of a protein chain. For this, we ran simulations of 12-residue-long polyalanines with and without the global sterics. The resulting

Table 6. Optimized Values of the Center-to-Center Distances between the Different Pairs of van-der-Waals Spheres for the Global Steric Interactions[a]

	CO	NH	C_α	GLY-C_α	P-C_β	PRO-C_β	H-C_β
CO	0.36	0.28	0.3	0.3	0.36	0.43	0.36
NH		0.34	0.34	0.34	0.4	0.4	0.4
C_α			0.32	0.32	0.36	0.38	0.36
GLY-C_α				0.4	0.44	0.44	0.44
P-C_β					0.4	0.44	0.4
PRO-C_β						0.44	0.44
H-C_β							0.4

[a]The distances are given in nanometers. For the N-terminal N, the same parameters as for the peptide−NH are used. Whereas for the polar side chains (P−C_β) the same averaged sphere position is used, the positions of the C_β differ for each of the hydrophobic side chains (H−C_β), though the radius is the same. The proline-side-chain sphere (PRO−C_β) was determined from the planar configuration, which is about the average between the up and down configurations.

radial distance distributions are given in Figure 6. The peak between 0.3 and 0.4 nm comes from the neighboring C_α's along

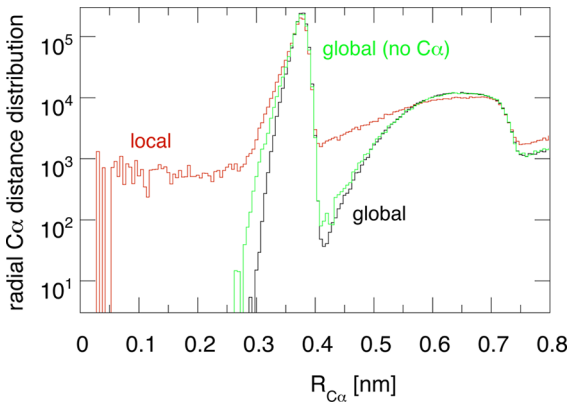

Figure 6. Effect of the global steric exclusions on the overlap of nonadjacent residues. From simulations of a 12-residue-long polyalanine, the distances between the C_α were extracted when either only local sterics (red curve, "local") or local plus global clashes (black curve, "global") were used. Without the global steric exclusions, non-neighboring residues could overlap, as can be seen from the nonzero probability for two C_α's to get closer than the next-residue distance of ≥0.3 nm. Ignoring the C_α (green curve, "global (no Cα)") considerably reduces the computational costs but has only minor effects on the radial C_α distance distribution.

the protein chain, and the broad hump around 0.6 to 0.7 nm is from the next neighbors. These two features are rather independent of the global sterics. When there are no global steric interactions (red data, labeled "local"), the protein chain may overlap with itself as can be seen from the nonvanishing density for distances below 0.3 nm and between 0.4 and 0.5 nm. When the C_α sphere is omitted (green data, labeled "global (no Cα)"), the chain does also not overlap with itself, and the resulting radial density distribution is only slightly widened. This shows that the deeply buried C_α is not very important for the spatial shape of the protein chain and may potentially be omitted to save computation time. However, we will postpone this issue until the model is complete with all interactions. Then, we can see how large the actual savings are when the C_α is omitted and whether it affects, e.g., secondary structure conformation or stability.

Extended Local Sterics. Beyond the local steric clashes that shape the dihedral distributions of the individual residues and the global sterics that determine the compactness of the folded protein chain, there are some more steric clashes between pairs of atoms which add a first bias for secondary structure formation. In α-helices, for example, non-neighboring residues separated by one turn may come so close that either their backbone atoms touch or the bulky side chains of the aromatic and β-branched residues collide with the backbone, which has a helix-breaking effect. In this example, the secondary structure neighborhood affects the conformation of an adjacent residue.

To identify such neighborhood effects, we used high-resolution structures of α and 3_{10}-helices from our *side chain set* reference data and extracted the dihedral angle distributions of the first and the second residues before and after the helices. These were compared to distributions for dipeptides, where there are no such secondary structure constraints. We found that residues directly following an α-helix have an increased probability to be in an α-configuration, too. To identify the respective atom−atom collisions that lead to this behavior, the last five residues of a helix plus the first nonhelix residue were extracted from 500 structures with an α-helix. From each residue, the side chain atoms beyond the C_β were removed. Then, the five helix residues were fixed in space, and the last residue, which directly follows the helix, was rotated around the φ and ψ angles. An overlap value $\varepsilon(\varphi,\psi)$ was computed for every pair of atoms at each (φ,ψ) grid point from the sum of the van-der-Waals radii of the atoms, r_1 and r_2, and the mutual distance d_{12} as

$$\varepsilon(\varphi, \psi) = \max\left(\frac{r_1 + r_2}{d_{12}}, 1.0\right) \tag{7}$$

The overlap value $\varepsilon(\varphi,\psi)$ was bounded below to 1.0 to make sure that rare but severe clashes were not lost during the subsequent averaging over the 500 individual structures. From these clash maps, we found that the most relevant overlaps occurred between the oxygens of residues i and $i-3$ and $i-4$, respectively, when the dihedral angles of the rotated residue i were in the β region. This increases the probability that the residue following an α-helix will also be in an α configuration and thus promotes growth of helical regions. These extended local steric collisions were implemented in our model analogously to the local sterics by van-der-Waals bonds between the respective positions of the involved pairs of oxygen atoms.

Weaker clashes were also found between the H and N of the peptide bond and the oxygens three and four residues away, but these only led to shifts within the α region. They do not alter the secondary structure propensities and were therefore not implemented in the model. A similar analysis was performed for β sheets and turns, too, but for these we found no steric clashes that would bias the conformation of adjacent residues.

A complementary helix-breaking effect comes from the bulky side chains of the β-branched residues isoleucine, valine, and threonine and the aromatic residues tryptophane, phenylalanine, and tyrosine. To identify the relevant clashes, another 500 structures were extracted from the *side chain set* reference data in which five residues in a helical conformation were followed by a β-branched residue. From the side chain of the β-branched residue the C_β and the two heavy atoms bound to it were kept, while from the five helix residues the side chain

dx.doi.org/10.1021/ct3005529 | J. Chem. Theory Comput. 2012, 8, 4732−4745

beyond the C_β was removed. Then a similar averaged clash value (see eq 7) was calculated when the first side chain dihedral χ_1 of the β-branched residue was varied. Again, clashes were identified between the $C_{\gamma1}$ and $C_{\gamma2}$ side chain atoms (or the respective $O_{\gamma1}$ in the case of threonine) and the oxygens three and four residues earlier, i.e., from the preceding turn of the helix. These clashes were also implemented as van-der-Waals bonds between the peptide bond oxygens O_{i-3} and O_{i-4} and the C_γ's (O_γ) of the β-branched residue i. The helix-breaking effect of the bulky side chains of the aromatic residues tryptophane, phenylalanine, and tyrosine was implemented in the same way as for valine, i.e., any further differences between their side chains are ignored at the current stage of the model.

The parameters of the van-der-Waals bonds used to model these extended local steric clashes are summarized in Table 7.

Table 7. Extended Local and β-Branching Atom−Atom Collisions Considered in Our Model[a]

atom pair	R_0 [nm]	remarks
$O_{(i-3)}-O_i$	0.360	
$O_{(i-4)}-O_i$	0.360	
$O_{(i-3)}-C_{\gamma1/2i}$	0.345	two bonds, for ILE, VAL, TRP, PHE, TYR
$O_{(i-4)}-C_{\gamma1/2i}$	0.345	-″-
$O_{(i-3)}-O_{\gamma i}$	0.340	for THR
$O_{(i-3)}-C_{\gamma i}$	0.345	-″-
$O_{(i-4)}-O_{\gamma i}$	0.340	-″-
$O_{(i-4)}-C_{\gamma i}$	0.345	-″-

[a]The bulky side chains of the aromatic residues TRP, PHE, and TYR are implemented as VAL concerning their helix-breaking sterics. For all these van-der-Waals bonds, C_{12} = 0.6 kJ mol^{-1} and dr = 0 were used (see eq 1).

How the extended local sterics and the β-branching clashes affect the dihedral angle distributions and thus the secondary structure propensities of the respective residues is illustrated in Figure 7. It gives the dihedral angle distributions from simulations of five alanines followed by a valine residue. The five alanines were fixed in an α-helical configuration to resemble the last turn of a helix. Figure 7A shows the observed dihedral distribution of the valine residue when only the local sterics of a generic residue were used. This is the same set of steric clashes as in a dipeptide where there is no secondary structure. Adding the extended local sterics (panel B) did not affect the α region but reduced the (already low) probability to find the valine in a β-conformation. The clashes with the β-branched side chain had the opposite effect of increasing the density in the β region. Also, φ was shifted to slightly more negative values (panel C). The combined effects of the extended local sterics and the β-branching collisions (panel D) do not greatly alter the dihedral distribution compared to panel A, which can be interpreted such that the β-branched collisions compensate the helix promoting effect of the $O_i-O_{i-3/4}$ clashes such that the next residue may be in any conformation.

Polyalanine Polymers. All the nonbonded interactions described so far are repulsive. Thus, no stable structures will fold, and a model protein should essentially behave like a polymer with a restricted local flexibility. For a bead−spring polymer, analytic scaling relations exist which, e.g., relate the average distance between the first and the last bead or the average radius of gyration to the chain length.[52,53] These static

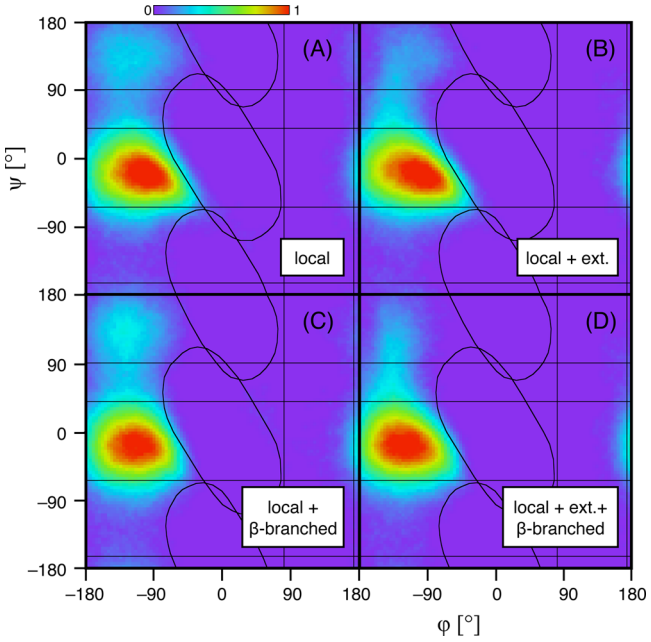

Figure 7. Effect of extended local steric exclusions and beta-branching on the dihedral distribution of a valine residue following an α-helix. In panel A, only local steric exclusions were used in the simulations. When extended local sterics were added (panel B), the probability in the β region was reduced, whereas β-branching shifted the probabilities toward more negative φ values (panel C). Panel D finally shows the combined effects of extended local steric and β-branching clashes. For these plots again a square root scaling was used.

properties, derived from self-avoiding random walks, are independent of the absolute values of the diffusion coefficients and also of the hydrodynamic interactions. The scaling relations predict that for a polymer with N beads the averaged squared radius of gyration scales as $\langle R_g^2 \rangle \propto (N-1)^{2\nu}$ with $2\nu = 1.16$.

To check whether our protein model with the bonds and steric clashes implemented so far behaves like a simple polymer with purely repulsive interactions between the beads, we ran simulations of polyalanines of N = 4...24 residues, i.e., for 10...50 building blocks, and for different levels of included interactions. The most basic setup only contained the bonds between the building blocks and the angle stabilizing bonds. It therefore had the same conformational flexibility as a protein backbone but without any restrictions on the dihedral angles. The averaged R_g^2 obtained from this most simple setup is labeled "bonds" in Figure 8, which gives the results from the simulations. As can be seen, the resulting $\langle R_g^2 \rangle$ scales as $N^{1.16}$. The results did not change when the local shapes were added, i.e., when the dihedral angles were confined to the allowed regions (data points labeled "local" in Figure 8). This indicates that protein chains with a length of more than a few residues essentially behave like a polymer on a global scale, even though the local dynamics may be different. When additionally the globally visible steric shapes were added, the scaling again did not change; only the average diameter of the coiled protein increased by about 50% ("global" data in Figure 8). As the "all sterics" data show, adding the extended local O−O interactions only marginally increased the diameter of the protein coil any further.

A first estimate of the numerical performance of our coarse-grained protein model can be obtained from the runtimes plotted in Figure 9, which gives the wall clock times for 1

dx.doi.org/10.1021/ct3005529 | J. Chem. Theory Comput. 2012, 8, 4732−4745

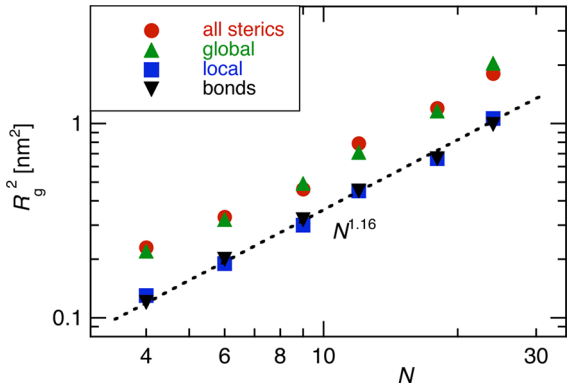

Figure 8. Average radius of gyration R_g^2 of polyalanine chains vs their number of residues, N, with the various hierarchies of steric interactions. The broken line gives the theoretically predicted scaling for a bead−spring polymer. In the most simple case ("bonds"), the peptide chain is modeled with only the bonds and no steric shapes at all. In the "local" case only the local sterics that limit the dihedral distributions were included. Beyond that, the "global" setup additionally contains the global steric clashes, whereas the extended local clashes were considered in the "all sterics" case, too. One sees that the global steric interactions that prevent a self-overlap are very important for the size of the protein, whereas the constraints on the dihedral angles from the local sterics and the extended local clashes have no visible influence at this averaged level.

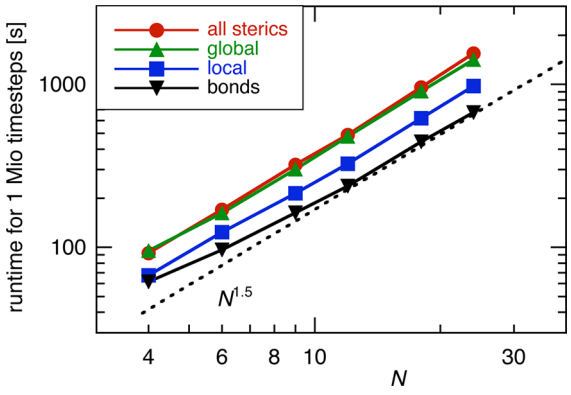

Figure 9. Runtime behavior of polyalanine chains vs their number of residues, N, for the different levels of steric interactions. The lines only connect the data points to serve as a guide to the eye. For these still rather short peptides, the runtime is mainly determined by the local interactions whose number grows roughly linearly with the chain length.

million simulation timesteps of polyalanine chains of various lengths, i.e., residue numbers N, and for the different levels of the steric interactions. One sees that for these still rather short peptides the runtimes scale roughly as $N^{1.5}$, i.e., faster than linear as one would expect for a bead−spring polymer with only next-neighbor interactions but still slower than quadratic when all interactions are globally visible. Interestingly, the full model with all steric interactions is only about 2 times slower than the bare model with only the bonds. From this, we can estimate that in these bare-bonds simulations less than one-third of the simulation time was required for the evaluation of the bonded forces, while the rest was needed for actually propagating the beads and the necessary book-keeping.

These simulations were performed with a conservative time step of Δt = 6 fs, i.e., the total simulated time was 6 ns. However, tests indicated that the simulations remained stable even with 5-fold longer timesteps of 30 fs. With all interactions included, it took about half an hour on a single Core-2-Duo processor with a 2 GHz clock speed to simulate 1 million timesteps with a peptide of N = 24 residues. Consequently, already with the rather conservative Δt = 6 fs up to 300 ns of total simulation time can be accumulated per day for such a peptide. Assuming that the $N^{1.5}$ scaling is still observed for slightly longer peptides, some 130 ns of simulated time could for example be obtained per day for an Alzheimer-β peptide of 42 residues on a typical laptop computer. With a faster computer and longer timesteps, 1 μs per day is possible for Aβ42 on a single CPU. When comparing these numbers to typical all-atom molecular dynamics simulations, one should additionally keep in mind that our model uses an implicit solvent and that therefore the size of the simulation box does not affect the runtime. This is especially important for unfolded, extended configurations or when long-time diffusional properties are considered.

CONCLUSIONS

In this publication, we presented our ansatz for a coarse-grained model of the protein backbone. This model is built up hierarchically, starting from the atomistic structure and uses nonspherical building blocks. To describe the conformational flexibility of the backbone, two main units are needed, namely the planar peptide group and the C_α atom with the first part of the side chain. In addition to these two, small N- and C-terminal groups were included, too. Initially, each of these units is a shapeless container into which interaction sites of various types are placed at arbitrary positions. The interactions of the complete model include chemical bonds, angle-stabilizing bonds, steric clashes between atoms or groups of atoms, electrostatic and hydrophobic interactions, and hydrogen bonds.

Here, in this first part, we were concerned with the chemical bonds between the building blocks and various steric clashes. With only the bonds but no further shapes the Ramachandran dihedral angles are completely unconstrained. The purpose of the spatial shapes with their purely repulsive interactions is to limit this flexibility to within the experimentally observed regions. Here, already some residue-specific constraints that bias secondary structure propensities were implemented. All further non-shape-related interactions as electrostatics or hydrophobicity will be described in a subsequent publication. These interactions then further shift the dihedral distributions to include even more residue specific details.

There are three levels of spatial clashes in our model. The lowest level is the most important and also the most detailed. For these local atom−atom collisions, we followed an analysis by Ho et al.,[13,14] who refined the initial analysis by Ramachandran et al.[50] to a minimal set of colliding atom pairs that define the observed dihedral angle distributions. For a generic residue, for example, these are the O and N atoms of the peptide group and the first side chain atom, C_β. Interestingly, the C_α atom, which is the starting point for many other coarse-grained approaches, is not involved. This can easily be understood as the rotatable bonds are attached directly to the C_α, and thus its direct neighbors will always stay at the same relative positions independent of how these bonds are rotated. The three dihedral-shaping atoms are all located off-center on their respective building blocks, and thus in the model a few simple binary repulsions are sufficient to define the complex local conformational landscape of the backbone in a

dx.doi.org/10.1021/ct3005529 | J. Chem. Theory Comput. 2012, 8, 4732−4745

very natural and efficient way. No complex three- or four-body interactions or angle terms are required to define how the protein chain may fold on the scale of a few neighboring residues. This lowest, local hierarchy of interactions is therefore the foundation onto which the rest of the model is built, which then further de- and refines how the protein chain behaves on a more global level. In addition to this local structure, also a globally visible shape is required to, e.g., prevent the protein from overlapping with itself. As the local details are already well-defined by the atom−atom clashes, the global shapes can be defined rather crudely. For the peptide bond a single sphere is obviously too simplified to capture its planar shape. We therefore used two spheres for a dumbbell-like shape. The C_α building block was modeled with one sphere on the C_α and one on the C_β, which already describes a part of the side chain. Test simulations, however, showed that the C_α sphere may potentially be omitted, as it is buried between the C_β sphere and the two adjacent peptide units, leaving us with only three (off-center) spheres per residue. Again, a small number of binary interactions is enough to describe a complex flexible shape.

With the local steric clashes the protein backbone is essentially finished, and any further interactions can only modify the probabilities to fold into one or the other secondary or tertiary structure motive. Here we described two such biasing interactions. The first are extended local clashes between the oxygen atoms three and four residues apart. These clashes support helix formation because a residue at the end of a helical segment is pushed toward an α-helical conformation. An opposite, helix-breaking effect comes from clashes of the peptide oxygens of the previous helix turn with the bulky side chains of β-branched or aromatic residues. Both effects could be added easily to the model with a few repulsive binary interactions.

At its current level the backbone model consists, apart from the chemical bonds that connect the building blocks, of repulsive interactions only. Therefore, secondary structure elements are not stable yet, and the model behaves like a polymer with a slightly more complex local flexibility as demonstrated with simulations of polyalanine chains of various lengths. From these simulations, we can already obtain a first estimate of the scaling of the runtime with the system size and the numerical performance. Without any parallelization, it will for example be possible to accumulate up to a microsecond of simulation time for a protein of the size of the Alzheimer Aβ42 peptide on a recent computer. It should also be noted that in our ansatz the propagation is based on the microscopic view of diffusion, which means that time scales are reproduced correctly when realistic diffusion coefficients are supplied for each of the building blocks.[15] Thus, there is no speedup even though the energy landscape is simplified and smoother than in the corresponding atomistic representation. With this "macroscopic" propagation scheme, we are thus able to mix in even more coarse-grained representations of other proteins, membranes, or vesicles without losing consistency. In a scenario, where for example the concurrent binding and folding of a peptide to a larger stable protein is investigated, the peptide would be modeled in detail, while the folded protein could be described as one rigid body with the respective surface shape and charge, saving the computational time for its internal dynamics.

With our hierarchic approach we thus presented a coarse-grained model of the protein backbone that is detailed on the local level but still efficient so that it can be used even for large systems. The hierarchic design allows one to parametrize the different interactions one after the other. Each of the interactions can also be switched on or off independently so that one can verify how much a given behavior depends on each of the interactions. The layered, hierarchic design which treats different physical potentials independently, furthermore allows one to modify or exchange any of the interactions in a future update without the need to reparameterize the complete model.

As said above, up to here only the concept and the first steric layers have been described. These allow the formation of secondary structures but do not stabilize them yet. The attractive interactions required for this will be introduced in a subsequent publication. With these attractions added, stable secondary structures can fold in a residue-specific manner. Then the model can be used to study for example the folding of small peptides or protein fragments or the dynamics of intrinsically disordered proteins. For the folding of larger globular proteins, the currently crudely simplified side chain representation has to be extended in a similar fashion as for the backbone. First the relevant rigid parts of each side chain are identified, and then the conformationally accessible rotamers are defined via a small number of local steric constraints. On top of this scaffold, simplified spatial representations, hydrophobic patches, and some point charges will complete the side chain representations and allow one to study the complete range of scenarios.

AUTHOR INFORMATION

Corresponding Author

*E-mail: andreas.wagenmann@googlemail.com; tihamer.geyer@bioinformatik.uni-saarland.de.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Uversky, V. N. *Curr. Alzheimer Res.* **2008**, *5*, 260−287.
(2) Kayed, R.; Head, E.; Thompson, J. L.; McIntire, T. M.; Milton, S. C.; Cotman, C. W.; Glabe, C. G. *Science* **2003**, *300*, 486−489.
(3) Shaw, D. E. *Proceedings of the ACM/IEEE Conference on Supercomputing, Networking, Storage and Analysis (SC09)*; ACM: Portland, OR, 2009.
(4) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Proc. Natl. Acad. Sci U. S. A.* **2012**, published ahead of print July 20, 2012.
(5) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102−106.
(6) Wolff, K.; Vendruscolo, M.; Porto, M. *PMC Biophys.* **2008**, *1*, 5.
(7) Cossio, P.; Marinelli, F.; Laio, A.; Pietrucci, F. *J. Phys. Chem. B* **2010**, *114*, 3259−3265.
(8) Bellesia, G.; Jewett, A. I.; Shea, J.-E. *Protein Sci.* **2011**, *20*, 818−826.
(9) Liwo, A.; He, Y.; Scheraga, H. A. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16890−16901.
(10) Pellarin, R.; Guarnera, E.; Caflisch, A. *J. Mol. Biol.* **2007**, *374*, 917−924.
(11) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10−15.
(12) Marring, S. J.; Risselada, H. J.; Yefimov, S.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812−7824.
(13) Ho, B. K.; Thomas, A.; Brasseur, R. *Protein Sci.* **2003**, *12*, 2508−2522.
(14) Ho, B. K.; Brasseur, R. *BMC Struct. Biol.* **2005**, *5*, 14.
(15) Winter, U.; Geyer, T. *J. Chem. Phys.* **2009**, *131*, 104102.
(16) Frembgen-Kesner, T.; Elcock, A. H. *J. Chem. Theory Comput.* **2009**, *5*, 242−256.

(17) Frembgen-Kesner, T.; Elcock, A. H. *Biophys. J.* **2010**, *99*, L75−L77.
(18) Geyer, T. *BMC Biophys.* **2011**, *4*, 7.
(19) Khalili, M.; Liwo, A.; Rakowski, F.; Grochowski, P.; Scheraga, H. A. *J. Phys. Chem. B* **2005**, *109*, 13785−13797.
(20) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 134105.
(21) Bereau, T.; Deserno, M. *J. Chem. Phys.* **2009**, *130*, 235106.
(22) Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(23) Maisuradze, G. G.; Senet, P.; Czaplewski, C.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. A* **2010**, *114*, 4471−4485.
(24) Wei, D.; Wang, F. *J. Chem. Phys.* **2010**, *133*, 084101.
(25) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623−644.
(26) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.
(27) Hills, R. D.; Lu, L.; Voth, G. A. *PLoS Comp. Biol.* **2010**, *6*, e1000827.
(28) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47−L49.
(29) Daggett, V.; Fersht, A. R. *Trends Biochem. Sci.* **2003**, *28*, 18−25.
(30) Wolff, K.; Vedruscolo, M.; Porto, M. *Phys. Rev. E* **2011**, *84*, 041934.
(31) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **1999**, *110*, 11616−11629.
(32) Takada, S. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 85−98.
(33) Fujitsuka, Y.; Takada, S.; Luthey-Schulten, Z. A.; Wolynes, P. G. *Proteins: Struct., Funct., Genet.* **2004**, *54*, 88−103.
(34) Smith, A.; Hall, C. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 344−360.
(35) Urbanc, B.; Cruz, L.; Yun, S.; Buldyrev, S.; Bitan, G.; Teplow, D.; Stanley, H. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17345−17350.
(36) Ding, F.; Borreguero, J. M.; Buldyrey, S. V.; Stanley, H. E.; Dokholyan, N. V. *Proteins: Struct., Funct., Genet.* **2003**, *53*, 220−228.
(37) Mu, Y.; Gao, Y. Q. J. *Chem. Phys.* **2007**, *127*, 105102.
(38) Forcellino, F.; Derreumaux, P. *Proteins: Struct., Funct., Genet.* **2001**, *45*, 159−166.
(39) Irbäck, A.; Sjunnesson, F.; Wallin, S. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *95*, 13614−13618.
(40) Favrin, G.; Irbäck, A.; Wallin, S. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 99−105.
(41) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.
(42) Schulz, J. C. F.; Schmidt, L.; Best, R. B.; Dzubiella, J.; Netz, R. R. J. *Am. Chem. Soc* **2012**, *134*, 6273−6279.
(43) Geyer, T.; Gorba, C.; Helms, V. *J. Chem. Phys.* **2004**, *120*, 4573−4580.
(44) Geyer, T.; Winter, U. *J. Chem. Phys.* **2009**, *130*, 114905.
(45) Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69*, 1352−1360.
(46) Einstein, A. *Ann. Phys.* **1905**, *17*, 549−560.
(47) O'Boyle, N. M.; Banck, M.; Craig, J. A.; Morley, C.; Vandermeersch, T.; Hutchinson, G. R. *J. Chemoinf.* **2011**, *3*, 33.
(48) Frishman, D.; Argos, P. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 566−579.
(49) Engh, R. A.; Huber, R. *Acta Crystallogr., Sect. A* **1991**, *47*, 392−400.
(50) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *J. Mol. Biol.* **1963**, *7*, 95−99.
(51) Word, M. J.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. J. *Mol. Biol.* **1999**, *285*, 1711−1733.
(52) Li, B.; Madras, N.; Sokal, A. D. *J. Stat. Phys.* **1995**, *80*, 661−754.
(53) Liu, B.; Dünweg, B. *J. Chem. Phys.* **2003**, *118*, 8061−8072.