Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Molecular Graphics and Modelling 22 (2004) 405–413

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Development of softcore potential functions for overcoming steric barriers in molecular dynamics simulations

Viktor Hornak [a,*], Carlos Simmerling [a,b]

[a] Center for Structural Biology, Stony Brook University, Stony Brook, NY 11794, USA
[b] Department of Chemistry, Stony Brook University, Stony Brook, NY 11794, USA

Abstract

In this work, we describe the development of softcore potential functions that permit occasional "tunneling" through the regions of conformational space during molecular dynamics (MD) simulations, which would otherwise be sterically prohibited. The modification consists of a truncation of the nonbonded interaction before the steeply repulsive region encountered at short interatomic distances. This modification affects both Lennard-Jones and Coulomb parts of the nonbonded potential. Critical to success is the choice of appropriate pairwise switching distances at which this modification should be made. In the present work, these are calculated based on potential of mean force functions extracted from model system molecular dynamics simulations. We believe that these functions describe the dynamic short-range interactions much better than mean force potentials derived from an ensemble of static structures (e.g. protein data bank (PDB)). Once a set of mean force potentials is obtained, a single empirical parameter, effective barrier height, is employed to determine switching distances for all pairwise atomic interactions. Changing this single parameter allows adjustment of the "softness" of the whole system. We tested the applicability of the new softcore potentials in a loop structure optimization study. The H1 loop in the antibody 17/9 was selected as our test case because substantial repacking of loop residues in the dense protein environment is necessary for successful relaxation of random initial conformations. Softcore simulations converted to correct loop conformations, in contrast to standard simulations which never sampled this structure even after 10 ns. The resulting root mean square deviation (RMSD) values (below 1.3 A for all heavy atoms of the loop) demonstrate the usefulness of the approach based on mean force derived softcore functions.

Keywords: Softcore potential; Potential of mean force; PMF; Steric barriers; Conformational sampling; Loop prediction; Molecular dynamics

1. Introduction

Efficient sampling of conformational space is an important aspect of biomolecular modeling. Many different methodologies have been developed which attempt to address the various difficulties that prevent good sampling. In cases where infinite steric barriers impose limitations on sampling which cannot be overcome by traditional techniques, such as simulated annealing, methods based on modification of steep repulsive barriers of atom pair potentials are often employed. The modified pair potentials are then usually referred to as softcore potentials. The first approach of this kind was proposed by Levitt [1] who used a modified van der Waals potential with restraints based on experimental data to obtain native-like folded conformations of a protein starting from random structures. Similar techniques were later used in NMR structure refinement with distance restraints [2,3] where the nonbonded interaction was only represented by a soft repulsive term; no electrostatic contribution nor attractive Lennard-Jones terms were included in this potential function. This may work well for refinement where substantial experimentally-derived distance restraints are present, but is not suitable for typical unrestrained molecular dynamics (MD) simulations. Softcore potentials were also used for avoiding singularities in free energy calculations [4] by softening the Lennard-Jones potential energy function, and in smoothing of the potential surface coupled with the diffusion equation method during structure optimization [5]. Two other reports [6,7] presented the use of softcore potentials in the prediction of protein loop conformations. The repulsive part of the nonbonded pair potential was replaced by a power function calculated based on forcefield parameters for a given pair of interacting atoms. As we elaborate below, we believe that a better criterion for a switching distance can be obtained from effective interactions rather than individual pair interactions.

The focus of this study is on conformational sampling, with the assumption that the underlying energy function is sufficiently accurate. Our objective was to formulate softcore

* Corresponding author. Tel.: +1-631-632-1439.
E-mail address: Viktor.Hornak@stonybrook.edu (V. Hornak).

1093-3263/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2003.12.007

potentials which would be used for structure optimization and refinement using molecular dynamics simulations with no need for additional restraints. We desire occasional "tunneling" events, in which groups of atoms might pass through others. These events may provide shortcuts through conformational space, permitting a more rapid transition than would be obtained by taking a more physically relevant path. For example, repacking residues in a protein core may normally involve partial unfolding of the protein, a slow process which can be avoided with the softcore potential. Thus, the resulting simulations are more appropriate for structure optimization than for studying actual transition paths.

In choosing distances for which tunneling is permitted, it is important to distinguish between individual pair potentials, as specified in the forcefield, and actual *effective* potentials, which arise from the complex interplay of all of the terms in the force field and include contributions from atoms not in the pair being considered. These effective potentials reflect the interactions that determine the net forces on atoms during typical molecular dynamics simulation of complex systems, such as proteins. We thus, focus on obtaining reliable switching distances which, contrary to previously described methods, are not based on properties of an interacting pair of isolated atoms (represented by the forcefield potential). In other words, we do not choose switching distances solely from the pair potential for the atoms involved in the tunneling event. Instead, we rely on statistical properties extracted from an ensemble obtained from model system MD simulations. Since we are interested in the behavior of atoms at short (contact) distances, the model system is constructed in a way which mimics the typical short-range interactions sampled in simulations of larger biomolecular systems, yet also includes more complex contributions that are not apparent from the pair potential.

One way to obtain effective potentials is to look at large numbers of experimentally determined native structures. There have been many studies [8–12] that extract such "knowledge-based" potentials from structures in the protein data bank (PDB). The frequencies of observation of particular structural features are then converted into free energies, or potentials of mean force (PMFs) [13,14]. These potentials are assumed to contain information about the overall folds of proteins (or protein–ligand interactions) and can, therefore, be used as scoring functions in structure prediction and drug design studies.

In contrast, our focus rests on a different aspect of structural information that is more dynamic and short-ranged in nature. Experimental structures represent low-energy or average conformations, and are poorly suited to extracting detailed contact effective energy profiles. We therefore, use ensembles of structures obtained from MD simulations, which provide the fluctuations in atomic contacts that are required for determining well-converged potentials of mean force at short distances. A set of PMFs are derived, with one for each pair of atom types, which we describe below. This approach also ensures that the PMF profiles are consistent with the force field that is employed during subsequent MD simulations.

The resulting effective potentials are then employed during structure optimization MD simulations. It is important to note, however, that we still employ the standard force field potentials to calculate atomic forces during MD; the effective potentials are only used to determine distances at which tunneling events are permitted. Since the effective potentials describe the (free) energy required to obtain a particular contact distance, it is possible to assign a single effective energy barrier to tunneling that can be simultaneously applied to all atom pairs. The size of this value then determines how frequently the tunneling events occur during the MD simulation.

We demonstrate the usability of these effective potentials on a successful optimization of a CDR loop in the antibody 17/9 [15]. We previously described the principal points of the loop optimization methodology [16]. The important difference lies in the present use of PMF-based switching distances. In our previous study, we had calculated interatomic distance histograms, and used these to select a switching distances corresponding to an empirical fraction of the histogram area below the sum of van der Waals radii for the pair. Our use of energy based PMFs, instead of the physically less meaningful histogram criteria, simplifies and rationalizes the selection of an effective barrier height at which a given pair switches to the softcore regime. Another important difference resides in our effort to make the PMF functions more transferable by introducing a non-protein model system rather than using specific protein systems. The latter approach may limit the applicability of the derived PMFs for optimization of conformations in other protein systems.

2. Methodology

2.1. Obtaining distance distribution functions

A typical MD simulation for a biomolecular system may involve a protein with the initial structure taken from X-ray or NMR experiments. After a short equilibration phase, an initial experimental structure would relax according to the forcefield used and start to explore conformations in the local energy basin. How far it ventures during a given timescale depends on factors such as the simulation temperature. At typical temperatures around 300 K, one usually only sees the structure fluctuate around its local energy minimum during nanosecond-length simulations, yet these fluctuations contain important information about the flexibility of individual atoms in their well-packed protein environment. If we look at how close different nonbonded atom pairs get to each other, we should see that, depending on their atom types, certain distances are preferred more than others. The shortest of such distances would usually reflect a contact interaction of the atom pair where the dynamic aspect, manifested through fluctuations of atoms, is taken into account.

Even though the interaction of the two atoms is described by the particular molecular mechanics forcefield, the contact distances sampled during MD do not necessarily correlate with distances that would be obtained from energy minimization with the corresponding forcefield potential function for the pair. That is not very surprising, because atoms in proteins are not isolated species and their interaction depends on a complex interplay of an observed atom pair with all other atoms in their environment. In other words, interactions mediated by nearby atoms will have an influence on how the two atoms "feel" each other. That is in fact reflected by the overall force on a particular atom calculated during MD simulation. A way to obtain more comprehensive information about the distances sampled by a given pair of atoms is to calculate the distance distribution function during a longer MD simulation. Unfortunately, the details of the distributions likely depend on the biomolecular system used. Each protein could produce distributions specific to its three-dimensional structure. The distances sampled by atom pairs may be rather restricted due to constraints imposed by the topology of the protein fold. Moreover, some atom types may not be represented in a given system and thus, no sampling of distances for such atoms could be obtained, or certain ranges may be missing due to small numbers of pairs. Such distributions may exhibit poor transferability from one protein system to another.

To avoid these problems we chose to perform the simulations on a model system which consists of free amino acid fragments, i.e. amino acids not terminated by amino or carboxyl groups, but rather ending with N–H and C=O groups (with unsatisfied valence states) identical to the ones they would have in a protein peptide bond. These fragments are placed into a simulation box and form, de facto, a system of liquid amino acids (a virtual "primordial soup"). The increased mobility of amino acids and lack of topological constraints allows sampling of many possible distances between atom pairs. Of course, by neglecting three-dimensional protein structure we are also likely to lose any information about secondary and tertiary structure in our distance distribution functions.

We carried out these amino acid simulations (as well as subsequent loop optimization studies) using the AMBER package [17]. Amino acid simulations were run for ~100 ns in a constant T, P ensemble at 300 K and 1 bar, with periodic boundary conditions and particle Mesh Ewald treatment of electrostatics [18]. The system contained 343 residues, where each of the 20 amino acids was represented approximately equally. The size of the simulation box after equilibration was about 31 × 53 × 35 Å. The forcefield used was ff94 [19]. Atoms were assigned to 16 groups of atom types, based on atom types in ff94 parameter set (Table 1).

Table 1
Atom types with their van der Waals radii (r) and description

Number	Name	r	Description
1	C	1.91	Carbonyl sp^2 carbon
2	CT	1.91	Any sp^3 carbon
3	C′	1.91	Aromatic, sp^2 double bonded, etc.
4	H	0.60	H attached to N
5	HO	0.60	H in alcohols and acids
6	HS	0.60	H attached to sulphur
7	H′	1.30	H to aromatic/aliphatic C
8	N	1.82	Amide N, sp^2 N in aromatic rings
9	N2	1.82	sp^2 N of aromatic amines
10	N3	1.88	sp^3 nitrogen
11	N′	1.82	sp^2 N in other rings
12	O	1.66	Carbonyl O
13	OH	1.72	Hydroxyl O
14	O2	1.66	Carboxyl O, ether O
15	S	2.00	SS sulfur in Met, Cys
16	SH	2.00	SH sulfur in Cys

In order to calculate distance distribution functions, we first constructed histograms of nonbonded interatomic distances collected during ~100 ns simulation; 1–2 (bond), 1–3 (angle), and 1–4 (dihedral) distances were excluded to avoid peaks in distance distributions due to angle and dihedral terms, and also because our focus is on nonbonded contact profiles. Distances were sampled in the range $0 < r < 14$ Å with a histogram bin width of 0.05 Å. The distance histograms were normalized to radial distance distribution functions:

$$g_{\mathrm{ab}}(r) = \frac{\Delta N_{\mathrm{ab}}(r)}{\rho_{\mathrm{b}} \Delta V(r)} \tag{1}$$

where $\Delta N_{\mathrm{ab}}(r)$ is the average number of b-sites located in the shell of radius r, and thickness Δr, centered on site a; $\rho_{\mathrm{b}} = N_{\mathrm{b}}/V$ is the average number density of b-sites in the system, where N_{b} is the number of b-sites in the system of volume V. ΔV is the volume of a thin shell at radius r, and thickness Δr. Sites a and b represent our atom types from Table 1. The denominator of Eq. (1) represents our "standard state," which is usually chosen to be a bulk density of b-sites in volume V. This ensures that as the distance between a- and b-sites increases, g_{ab} approaches unity, thus indicating that there is no structuring of a-sites and b-sites at longer distances. Such normalization is straightforward for our model system but is more problematic in the case of solvated protein systems, because water molecules would have to be excluded during calculation of the accessible volume shell ΔV, as well as during calculation of the average number density ρ_{b}.

It is convenient to transform distance distribution functions, $g_{\mathrm{ab}}(r)$, to free energy functions, also called potentials of mean force:

$$\mathrm{PMF}_{\mathrm{ab}} = -kT \ln g_{\mathrm{ab}}(r) \tag{2}$$

We determined a total of 136 (16 × 17/2) pairwise PMF functions for 16 groups of atoms. The values of these functions indicate the free energy change when two atoms are brought from infinity to a distance r. At short distances the function steeply rises, indicating that infinite free energy is required to overlap two atoms. This is the domain where we want to modify the pair potential.

The convergence of PMFs is dependent on a good sampling. The PMFs for atom types which are abundant in amino acids (e.g. C, O, CT) converge fairly quickly. However, less frequently occurring types (such as OH, S, SH) require substantial simulation times. To make sure that our converged PMFs were not affected by the initial configuration of our model system, we performed an independent 68 ns simulation where the composition of the system was the same but the initial positions of amino acid fragments were different. A comparison of the PMFs obtained from the two simulations showed that the functions were practically identical. As a measure of similarity, we calculated an average absolute difference between the alternate PMF functions over the range of distances where PMF values are less than $10kT$. All atom pairs had an average difference of their PMFs smaller than $0.23kT$ with the exception of the single largest value of $0.54kT$ for N2–N2 atom pair. However, typical differences are below $0.1kT$ (for 117 pairs out of all 136).

Another factor that we investigated was simulation temperature. The PMF functions should, in principle, depend on the temperature only through kT factor in Eq. (2). Therefore, we ran additional simulations of the same model system at two other temperatures 250 K (∼56 ns) and 350 K (∼62 ns) and confirmed that the PMF/kT functions extracted at these two temperatures are highly similar. For the 250 K simulation, there were 89 pairs with a less than $0.1kT$ difference from the 300 K data and 22 pairs within the range of 0.2–$0.58kT$. In case of 350 K simulation, there were 122 pairs with less than $0.1kT$ difference and five pairs with differences of 0.2–$0.32kT$. The similarity of these PMF functions suggests that it is reasonable to use the PMF potentials obtained at 300 K during MD simulations at other temperatures within that range.

In order to assess the effect of using only amino acid fragments to obtain the PMF curves, we derived comparable PMF curves for two proteins in explicit solvent. However, there is a difficulty with extracting pair distribution functions from these simulations. As discussed above, it is difficult to properly normalize these distance distribution functions in explicit solvent. When number densities of a particular atom pair are normalized by the volume of the accessible spherical shell, we need to account for the fact that not all of that volume is truly available because varying portions are occupied by solvent. Therefore, for the sake of comparison, we arbitrarily scaled PMFs obtained from the protein simulations to best fit the PMFs of our model system. Such scaling will at least allow qualitative comparison of the functions, and in particular the location of minima on the PMFs from all three systems.

2.2. Softcore simulations using the PMF data

Critical to the success of softcore simulations is a determination of switching distances at which the highly repulsive portion of the potential function will be replaced with a less repulsive form. In principle, each pair of atom types should have its own switching distance, located near the distance at which the atom pair starts to encounter a strong repulsive force. Such distances can be obtained from the PMF curves that we described in the previous section. The use of the PMF functions also permits specification of the effective barrier at the point of truncation of the potential. For example, one may obtain frequent tunneling events if the potential is truncated at an effective energy barrier of 1–$2kT$ on the PMF curve. We find that an efficient range of effective barriers is from about 2–$4kT$. Within that range, the number of tunneling events linearly decreases as we approach the upper limit because atoms require higher energies in order to be able to reach the higher energy switching distance and switch over to the softcore regime.

Each predefined value of a barrier (e.g. $3kT$) corresponds to a *set* of switching distances (one for each atom type pair) at which forcefield pair potentials switch to softcore pair potentials. This point is illustrated for a specific example in Section 3. The form of softcore potential that we employ is simple: the steeply repulsive part of the standard ff94 potential function is removed such that the normal nonbonded potential (Coulomb and Lennard-Jones terms) is replaced by a linear repulsive potential. This results in a constant repulsive force once the distance of the two interacting atoms drops below the switching distance r_{soft}:

$$\begin{aligned} r \geq r_{\text{soft}}: &\qquad E_{\text{nonbond}} = E_{\text{nonbond}}^{\text{AMBER}} \\ r < r_{\text{soft}}: &\qquad E_{\text{nonbond}} = -k(r - r_{\text{soft}}) + E_{\text{nonbond}}^{\text{AMBER}}(r_{\text{soft}}) \end{aligned} \tag{3}$$

Note that the energy, but not the force, is continuous during the softcore transition. In our previous study, we tested various switching functions and found no significant improvement in the behavior of the simulations [16].

The simulations were prepared as in our previous study and performed using a modified version of the AMBER suite of programs [17]. Our model system was the antibody 17/9. The goal of the simulation is to optimize the H1 CDR loop, which is seven residues long in 17/9 and contains both polar and hydrophobic amino acids with three bulky aromatic sidechains ($\text{Y}^{26}\text{SSFSFG}^{32}$). A 50-residue fragment of the antibody was used in the loop modeling simulations. Cartesian positional restraints (5 kcal/mol/A^2) were used to maintain the experimentally determined structure for the nonloop portion of the fragment. Solvent effects were approximated using a distance dependent dielectric constant, in the canonical ensemble at 300 K and with time step of 2 fs. The convergence of the simulated loop structures with respect to the X-ray structure was measured in terms of root mean square deviation (RMSD), with a best-fit of the non-loop regions followed by RMSD calculation for the loop heavy atoms. Thus, the RMSD measures the relative position of the loop and protein, in addition to the internal conformational variation of the loop. This is a more rigorous measure of deviation than is obtained if one performs a best-fit of the simulated and experimental (i.e. X-ray) loop regions.

Nine starting random structures of the H1 loop were generated via high-temperature (500 K) MD simulation. The initial heavy atom RMSDs with respect to the experimental structure of the loop ranged from 4.5 to 7.1 Å. The effective tunneling barrier height, which determines the distance at which the function is switched over to the softcore function, was periodically cycled between 2 and 4*kT* over 100 ps intervals. This was shown in our previous work to enhance the effect of softcore functions, and also helps to avoid the problem with choosing a specific value of this empirical parameter. In practice, when barrier height is around its lower limit (2*kT*) more atoms are likely to reach the barrier and switch to softcore regime, while when we approach the upper limit (4*kT*), very few pairs are capable of reaching the switching distance, and therefore we are effectively resuming unmodified forcefield potentials and restoring “physically reasonable” conformations. This procedure, thus, periodically destabilizes any new (unphysical) local energy minima that were introduced by the softcore function.

3. Results and discussion

In Fig. 1 we show examples of PMF functions for several atom type pairs. These particular pairs were chosen to point out certain characteristic features of the PMFs that are consistent with what one can intuitively infer from commonly observed hydrogen bond and charge–charge interactions.

The first graph (Fig. 1a) shows which distances are most likely to occur between three different pairs of atoms during our model system simulation. All three pairs are chosen from backbone atoms: amide nitrogen (N) and hydrogen (H), and carbonyl oxygen (O) and carbon (C). The first deep minimum in the H–O distance (∼1.9 Å) is a clear representation of hydrogen bond between carbonyl and amide backbone groups. The second minimum located over a range of ∼3.5–4.5 Å corresponds to distances between H and O involved in neighboring hydrogen bonds in α-helical secondary structures. This does not imply that such secondary structure elements are present in the simulation of our model system (recall that our simulation has amino acid fragments which are not connected and, therefore, cannot form a secondary or tertiary structure). It only indicates that, due to backbone hydrogen bonds, the fragments assume an orientation which might also be found in secondary structures. Minima on the other two PMF curves (N–O and C–N) in Fig. 1a are determined by the former hydrogen bond, which significantly restricts the number of possible orientations for these pairs.

The three plots in Fig. 1b show sampled distance preferences due to hydrogen bonding between hydroxyl groups of serine, threonine, and tyrosine. The first narrow minimum on the HO–OH curve again represents the direct hydrogen bond between the hydrogen acceptor group oxygen and the donor group hydrogen. The second broader minimum represents the distances sampled by the acceptor group hydrogen (which is not directly involved in the hydrogen bond) and the donor group oxygen. Simple geometry considerations indicate that these minima would indeed occur around 3.2 Å. The two minima on the HO–OH curve are also a good example of the observation that highly specific interactions display narrow shaped minima while distances which are indirectly induced or are not highly specific have broader minima. The other two plots in Fig. 1b are consistent with these conclusions.

The last example depicted in Fig. 1c points out another characteristic interaction in protein systems—ionic interactions (salt bridges) between charged amino acids, such as lysines, arginines, and glutamic and aspartic acids. Looking at several geometric configurations of their interacting amino and carboxyl groups, one can again rationalize most of the features found on the three plots in Fig. 1c. The first minimum at ∼2.7 Å between amino (or charged amino) nitrogen (N2, N3) and carboxyl oxygen (O2) is present due to direct charge–charge interaction. The other distinct minimum at a relatively long distance (∼4.9 A) can be explained

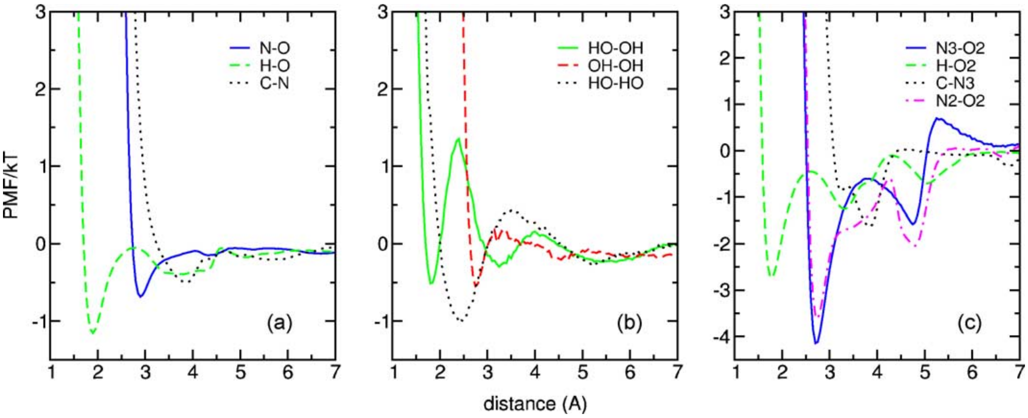

Fig. 1. The PMF functions of several atom pairs: (a) first case demonstrates the distance preferences following from hydrogen bonding of backbone carbonyl and amide groups; (b) another example of distance preferences induced by hydrogen bonding of hydroxyl groups; and (c) distances due to salt bridges in charged amino acids. For explanation of atomic types in legends see Table 1.

based on the geometric constraints due to the rigidity of the terminal guanidine group of arginines. The C–N3 and H–O2 curves can be explained on similar grounds. The more complex H–O2 curve results from the mixture of geometry constraints and hydrogen bonds of terminal amino groups of lysine and arginine to carboxyl oxygens of glutamic or aspartic acid.

In summary, the PMF curves may be viewed as a description of short and medium-range structuring factors. Specific interactions, like the ones described above, generally produce distinct minima at shorter distances, while nonspecific longer range or nonpolar interactions result in broader and shallower minima at longer distances (>4 Å).

Even though the PMF curves contain some interesting information about the effective interactions in our system, our present focus is directed at short-range properties. The first minima represent the most likely short-range contact distances between two atoms. If the atom pairs approach more closely than these values, the effective energy rises steeply. Therefore, the first minima are a good description of how close two atoms approach during MD simulation as a result of both the intrinsic pair potential and the effective multi-body interactions. The PMF functions corresponding to shorter distances than those minima give a quantitative measure of effective energy barriers that need to be overcome if the atoms are to move to such shorter distances.

For many pairs there is no obvious relationship between PMFs and pair potential functions calculated directly from the forcefield. This is not surprising, as the former characterizes effective interaction in a large and complex system, while the latter describes the interaction of an *isolated* pair of atoms. Fig. 2 contrasts forcefield potential functions and PMFs for four different atom pairs. On the left, carbonyl oxygen (O–O) and hydroxyl oxygen (OH–OH) pairs both have repulsive forcefield pair potentials over the whole range of distances shown. In contrast, their PMFs shows clear preferences for certain distances: ∼3.7 Å for carbonyl oxygens and ∼2.7 Å for hydroxyl oxygens. In addition, the short-range PMF curves demonstrate that OH–OH pairs sample much shorter distances than observed for O–O pairs, the reverse of what would be expected from the pair potential. The narrow minimum for the OH–OH pair corresponds to direct hydrogen bonding at distances much shorter than would be inferred from the pair potential. The broader minimum for O–O pairs arises from less specific interactions indirectly imposed by backbone (C=O · · · H–N) hydrogen bonding.

The other two pairs shown in Fig. 2, C–O and C–N, both have pair potentials with minima close to each other (∼2.7 Å for C–O and ∼2.9 Å for C–N), yet their PMFs show that the closest distances which these pairs preferentially sample are different: ∼3.1 Å for the C–O pair and ∼3.8 Å for the C–N pair.

We next performed a comparison of PMFs derived from our model system and the ones taken from simulations of fully solvated proteins (in contrast to amino acid fragments). We tested two explicitly solvated protein systems with different fold characteristics: (1) the full 17/9 antibody that we described above and (2) a small G-protein ras-p21 (pdb code 5P21). We carried out relatively long MD simulations of both systems: ∼18 ns for 17/9 and ∼20 ns for ras-p21. A few selected PMF functions for the two proteins together with the functions from our model system are shown in Fig. 3.

These examples show some of the typical patterns that emerge from the comparison. The upper two plots show

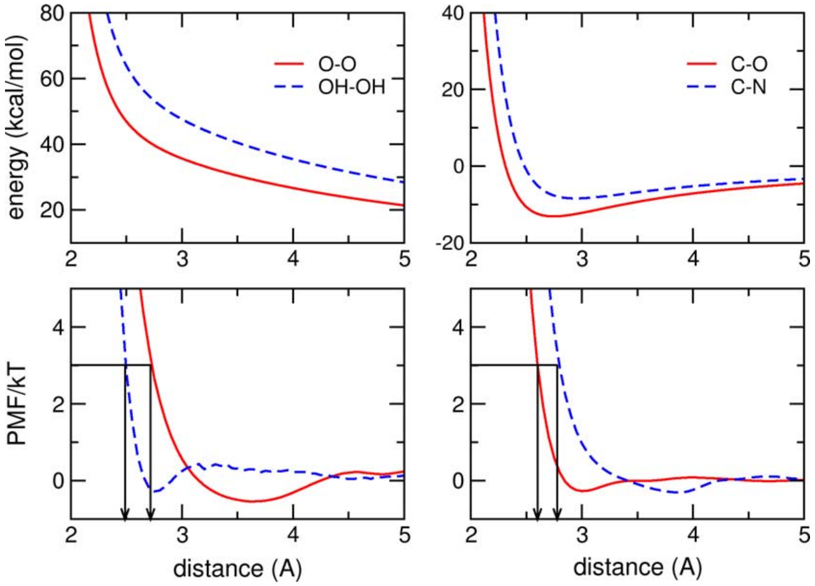

Fig. 2. The upper two plots show ff94 pair potentials (nonbond, i.e. Lennard-Jones and Coulomb) for four different atom pairs, first two (O–O and OH–OH) with repulsive forces over the whole distance range, other two (C–O and C–N) with minima. The corresponding PMFs in the bottom two plots show that the most probable distances at which these pairs are found can be quite different even in cases where pair potential functions are similar. Arrows indicate distances at which the PMF curves reach a barrier of 3*kT*.

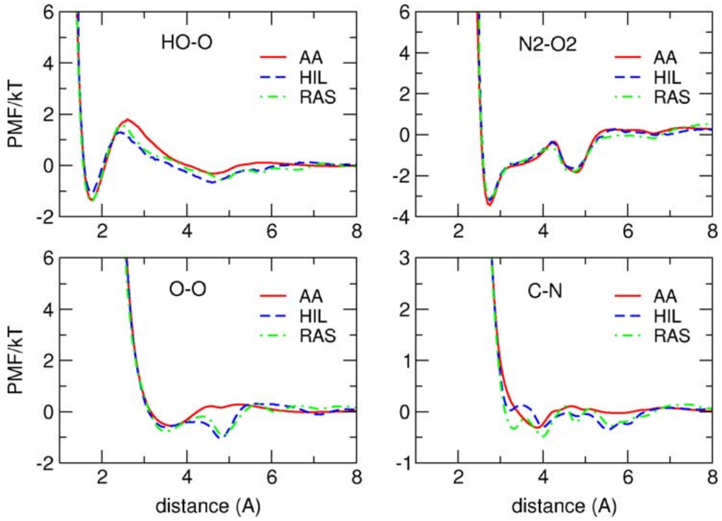

Fig. 3. Comparison of PMFs obtained from amino acid fragments (AA) and from two proteins (HIL—antibody 17/9, RAS—ras-p21). The upper two plots show atom pairs with very similar PMFs. The bottom two plots show PMFs which are similar to a certain extent but also have features that differ between proteins and amino acid fragments.

HO–O and N2–O2 functions, which are nearly identical. The bottom two plots show functions which exhibit some similarity, but there are features which distinguish protein PMFs from the model system functions. The minimum in the O–O protein curves at ∼4.8 Å (which is missing in the model system PMF) can be attributed to O–O distances in antiparallel β-sheets, as well as to one of the specific O–O distances in α-helices. In the other case, the C–N protein PMFs show two minima at ∼3.2 and ∼5.6 Å, which are not present in model system PMFs. The former minimum can be accounted for by C–N distances found in turns of antiparallel β-sheets as well as in (i, $i+2$) C–N distances in β-helices. The latter minimum can also be explained by C–N distances between two strands of β-sheets and (i, $i+3$) distances in helices.

In general, we see that the short-range PMF functions for given atom type pairs are very similar regardless of the system used to generate them: they show the same slope at short distances and frequently share the same position of the first minimum. This is reassuring, since this is the region where the PMF data will be employed for nonbonded force truncation and facilitation of tunneling. The differences at medium range distances are usually due to the presence of secondary structure elements. The details of the PMFs at longer distances usually differ, not only between proteins and our model system but also between the two proteins. This is likely caused by different three-dimensional topologies imposing different constraints on distance sampling. Also, many PMFs from protein systems suffer from insufficient sampling (or from the complete absence of certain atom types) and their comparison would, therefore, not be very meaningful.

In the next section, we show how we utilized the PMF functions derived from our model system for a typical application, in this case optimization of the conformation of a surface loop in a protein. We chose the same system as in our previous work [16] (the H1 loop of the antibody 17/9), where we also showed that normal MD techniques were not able to achieve correct native structures of the loop even after 10 ns MD simulation at 300 K (Fig. 4). All of the simulations are kinetically trapped near the region of the initial conformation.

The H1 loop of 17/9 is particularly suitable for testing softcore potentials because of the specific location of the loop sidechains with regard to the protein environment. Three sidechain aromatic rings, especially Phe27, are buried in pockets protruding under the surface of the protein, and the entire loop has relatively low solvent exposure (35% of the surface is solvent-accessible). Due to this extensive sidechain packing, the conformation of the H1 loop is very

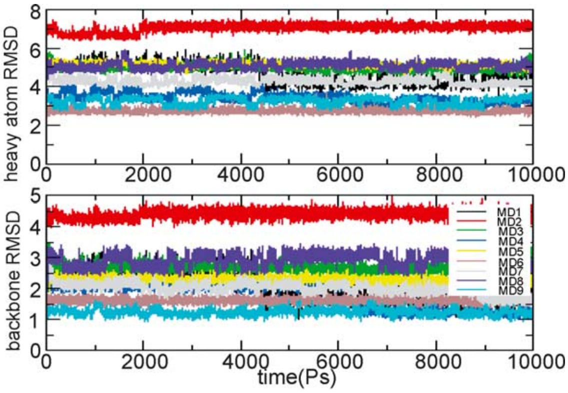

Fig. 4. Nine standard MD simulations (MD1-9) starting from the random initial loop structures. The heavy atom (top) and backbone (bottom) RMSDs with respect to the X-ray loop structure (fit to non-loop region) are shown. Simulations were run for 10 ns and we observe that none of the nine simulations was able to converge or even transiently sample the correct loop conformation on the 10 ns timescale.

well defined and exhibits low atomic positional fluctuations during the MD simulation of the native structure. This may not be the case for loops which are more solvent exposed, thereby resulting in increased loop flexibility and reduced structural definition, particularly for surface sidechain groups. In the initial random structures, none of the aromatics rings were packed into their corresponding pockets in the protein. To achieve the correct structure without partial unfolding of the remainder of the protein, the loop would have to find a very narrow path for packing the bulky aromatic sidechains. Our approach is to overcome this problem by allowing these groups to tunnel through the steric barriers to facilitate repacking of these side chains.

We ran nine independent simulations, each initiated from one of the random loop structures. The “softness” of the loop was controlled by specification of the (time-dependent) energy barrier in the input file. A particular value yields a set of switching distances (one for each atom type pair), which were consequently used by AMBER for the modification of forcefield potentials. The normal nonbonded potential was truncated for atom pairs whose distance during the simulation became less than their corresponding switching distance, according to Eq. (3). For example, if the energy barrier at a particular time step during the simulation was set to $3kT$, the particular pair of atoms would only switch to softcore potential if the effective energy from their corresponding PMF function becomes larger than $3kT$. More specifically, for an atom pair of type O–O, the nonbond forcefield potential switches to the softcore potential when the interatomic distance is less than 2.7 Å (see Fig. 2, left two plots). For another pair, OH–OH, the barrier of $3kT$ is achieved at a distance of 2.5 Å. It is important to note that this value is near the minimum in the PMF curve for the OH–OH pair, thus, necessitating accurate calculation of PMFs to avoid truncation in regions that are sampled in low-energy conformations.

With softcore potentials employed, in eight out of nine cases the correct loop structure is obtained (Fig. 5). About half of the simulations achieved the correct structure within 3 ns while the rest took much longer times to converge. One simulation did not yield the correct structure at all. A backbone RMSD plot of this simulation (the purple graph at the bottom plot of Fig. 5) indicates that the backbone is actually not far from its correct location and the small difference is found to be due to a backbone conformation around Phe27. A closer inspection of the trajectory and the initial random structure reveals that the peptide bond between Ser28 and Phe27 started in an incorrect *cis*-conformation which resulted from generating random structures by high-temperature MD. The softcore potential function does not, of course, improve sampling of such barriers.

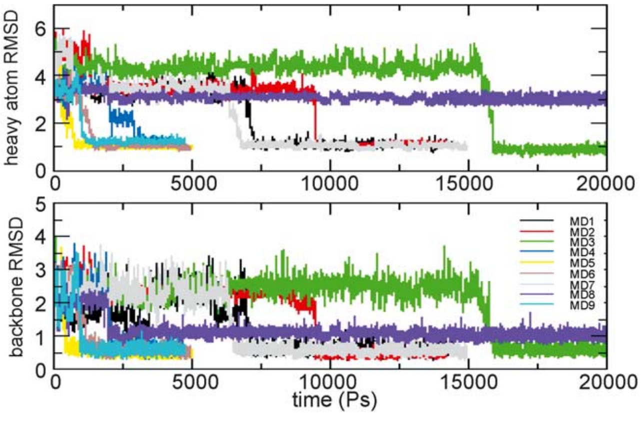

Fig. 5. Heavy atom (top) and backbone (bottom) RMSD plots of nine simulations (designated MD1–9) starting from random initial conformations. Eight structures out of nine converged to the correct X-ray structure. RMSDs of all structures (fit to nonloop region) that converged were below 1.27 Å for heavy atoms and below 0.68 Å for backbone atoms.

4. Conclusions

We obtained a set of atom type dependent PMFs from our “generic” model system that consisted of amino acid fragments. These PMF functions provide a detailed description of effective interactions of constituent units of proteins at short distances, on which we based a softcore modification for the nonbond portion of the potential energy function. The modified forcefield potentials (i.e. softcore potentials) are intended to be used in combination with MD simulations, as opposed to knowledge-based potentials which are mostly used as a discrimination tool (e.g. for conformational scoring).

We demonstrated that fairly comprehensive information about the short-range interactions in biomolecular systems can be inferred from the PMFs. In particular, we showed how typical hydrogen bonds and charge-charge interactions manifest themselves in the PMF functions. An important assumption about the transferability of PMFs obtained from our model (i.e. non-protein) system to real protein systems was examined. The qualitative comparison of the PMF functions from our model system and both protein systems was performed. PMFs obtained from full proteins differed from each other (and from the model system) in the longer-range details of the PMFs, a result of differences in protein topology as well as of a poor sampling.

As a sample application, this set of PMFs obtained from our model system were consequently used in enhancing the sampling of conformational space in a case where high steric barriers exist (which is fairly typical in condensed systems). We successfully applied the method for optimization of a loop conformation in a protein system. Simulations were able to locate highly native-like conformations for the loop within several nanoseconds, in contrast to standard simulations which never sampled native conformations on this timescale. Even though the basic methodology for the protein loop prediction was outlined in our previous article [16], an important generalization of distance preferences to energy-based PMF functions together with transferability of PMFs were introduced in this work.

Acknowledgements

Support for this project was provided by the AMDeC Foundation of New York City, through its “Tartikoff/Perelman/EIF fund for Young Investigators in Women’s Cancers,” and by NIH GM6167803. CS is a Research Corporation Cottrell Scholar. Portions of this work were carried out at the Pittsburgh Supercomputing Center under NPACI allocation MCA02N028.

References

[1] M. Levitt, Protein folding by restrained energy minimization and molecular dynamics, J. Mol. Biol. 170 (3) (1983) 723–764.

[2] M. Nilges, G.M. Clore, A.M. Gronenborn, Determination of 3-dimensional structures of proteins from interproton distance data by dynamical simulated annealing from a random array of atoms—circumventing problems associated with folding, FEBS Lett. 239 (1) (1988) 129–136.

[3] M. Ullner, et al., 3-Dimensional structure of the apo form of the N-terminal Egf-like module of blood-coagulation factor-X as determined by NMR-spectroscopy and simulated folding, Biochemistry 31 (26) (1992) 5974–5983.

[4] T.C. Beutler, et al., Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations, Chem. Phys. Lett. 222 (6) (1994) 529–539.

[5] T. Huber, A.E. Torda, W.F. vanGunsteren, Structure optimization combining soft-core interaction functions, the diffusion equation method, and molecular dynamics, J. Phys. Chem. A 101 (33) (1997) 5926–5930.

[6] K. Tappura, M. Lahtela-Kakkonen, O. Teleman, A new soft-core potential function for molecular dynamics applied to the prediction of protein loop conformations, J. Comput. Chem. 21 (5) (2000) 388–397.

[7] K. Tappura, Influence of rotational energy barriers to the conformational search of protein loops in molecular dynamics and ranking the conformations, Proteins-Struct. Funct. Genet. 44 (3) (2001) 167–179.

[8] J. Moult, Comparison of database potentials and molecular mechanics force fields, Curr. Opin. Struct. Biol. 7 (2) (1997) 194–199.

[9] F. Melo, E. Feytmans, Novel knowledge-based mean force potential at atomic level, J. Mol. Biol. 267 (1) (1997) 207–222.

[10] A. Rojnuckarin, S. Subramaniam, Knowledge-based interaction potentials for proteins, Proteins-Struct. Funct. Genet. 36 (1) (1999) 54–67.

[11] J.B.O. Mitchell, et al., BLEEP—potential of mean force describing protein-ligand interactions: I. generating potential, J. Comput. Chem. 20 (11) (1999) 1165–1176.

[12] I. Muegge, A knowledge-based scoring function for protein-ligand interactions: probing the reference state, Perspect. Drug Discovery Design 20 (1) (2000) 99–114.

[13] M.J. Sippl, Calculation of conformational ensembles from potentials of mean force—an approach to the knowledge-based prediction of local structures in globular-proteins, J. Mol. Biol. 213 (4) (1990) 859–883.

[14] M.J. Sippl, et al., Helmholtz free energies of atom pair interactions in proteins, Fold. Des. 1 (4) (1996) 289–298.

[15] U. Schulzegahmen, J.M. Rini, I.A. Wilson, Detailed analysis of the free and bound conformations of an antibody—X-ray structures of fab 17/9 and 3 different fab-peptide complexes, J. Mol. Biol. 234 (4) (1993) 1098–1118.

[16] V. Hornak, C. Simmerling, generation of accurate protein loop conformations through low-barrier molecular dynamics, Proteins: Struct., Funct., Genet. 51 (2003) 577–590.

[17] D.A. Case, D.A. Pearlman, J.W. Caldwell, T.E. Cheatham III, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, V. Tsui, R.J. Radmer, Y. Duan, J. Pitera, I. Massova, G.L. Seibel, U.C. Singh, P.K. Weiner, P.A. Kollman, AMBER, version 6, University of California, San Francisco, 1999.

[18] U. Essman, et al., A smooth particle Mesh Ewald method, J. Chem. Phys. 103 (1995) 8577–8593.

[19] W.D. Cornell, et al., A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules, J. Am. Chem. Soc. 117 (19) (1995) 5179–5197.