Reproducible Polypeptide Folding and Structure Prediction using Molecular Dynamics Simulations

M. Marvin Seibert1,2, Alexandra Patriksson2, Berk Hess3 and David van der Spoel2*

1Department of Mathematical Sciences, Chalmers University of Technology, SE41296 Gothenberg, Sweden
2Department of Cell and Molecular Biology, Uppsala University, Husargatan 3 Box 596, SE-751 24 Uppsala Sweden
3Max Planck Institute for Polymer Research Ackermannweg 10, D-55128 Mainz, Germany

*Corresponding author

0022-2836/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.

Introduction

Protein folding by molecular dynamics simulation has been termed the holy grail of computational biochemistry.1 Since the breakthrough paper on folding pathways of the villin headpiece2 the field of protein folding simulations has made great strides forward. For instance, the Pande group have introduced a new method, using distributed computing, to perform folding simulations and deduce folding kinetics from it.3–8 Other important contributions have been made by Simmerling et al.,9 who reported a folding simulation of the Trp-cage structure, and by Zhou, who has derived energy landscapes from replica exchange molecular dynamics (REMD) simulations.10–12 In addition, Garcia et al. have reported detailed studies of aspects of the protein folding, in particular the effects of pressure.13–17 Finally, the Van Gunsteren group have studied the folding thermodynamics of peptides composed of β-amino acid residues in detail,18–20 focusing in particular on the role of solvent in the entropy of folding.21

One record from the seminal Villin headpiece paper of Duan & Kollman2 still stands, that of the longest continuous MD simulation of a protein (1 μs), obtained using massively parallel computing on 256 processors. Here, we report very long classical trajectories (1.6 μs and 1.8 μs) and long (460 ns) replica exchange22 molecular dynamics simulations of the folding of chignolin, a β-hairpin peptide, in explicit water and in vacuo (1 μs REMD). The structure and folding characteristics of chignolin have been reported recently,23 based on nuclear magnetic resonance (NMR) experiments, allowing a detailed comparison of our simulations to experimental data.

Mini-proteins, protein fragments, and stable peptides have been studied for a long time as possible nuclei for protein folding.24–32 Similar structures are present in large proteins and, therefore, structure prediction of such peptides is an appropriate approach to the protein folding problem. It was shown previously that intricate interactions25 in short peptides25 can be reproduced by molecular dynamics (MD) simulations,33 also, that not all peptides have well-defined structures in

Abbreviations used: MD, molecular dynamics; REMD, replica exchange molecular dynamics; NOE, nuclear Overhauser effects; RMSD, root-mean-square deviation.

E-mail address of the corresponding author:
spoel@xray.bmc.uu.se

water, although they may have one in other environments, e.g. Leu-enkephalin.\textsuperscript{34} Obviously, the environment of a peptide matters. In particular, it has been shown that water can “lubricate” conformational transitions in proteins and peptides,\textsuperscript{35} and catalyze protein folding.\textsuperscript{36}

Here, we investigate the folding of chignolin in both water and \textit{in vacuo} using REMD simulations, starting always from extended configurations. The REMD technique\textsuperscript{22} allows for significantly enhanced sampling of conformations, due to frequent switching of simulation temperatures (based on a Metropolis criterion) between a number of concurrent simulations. In this case, 16 simulations with temperatures ranging from 275 K to 419 K were performed (see Methods), for a duration of 1 \textmu s \textit{in vacuo} and 460 ns in solvent. We show how the REMD algorithm can be used to systematically improve the structure of a peptide towards the correct native structure in solvent, even when starting from a random, extended state. The ensemble of structures with the lowest free energy (the most populated bin in an energy landscape analysis) corresponds to the native structure, consistent with Anfinsen’s dogma.\textsuperscript{37}

The vacuum simulations are not only interesting to highlight the role of water in protein folding, but are also relevant to the more general problem of the structure of a protein \textit{in vacuo}. This question is important in the context of biology in the gas phase, e.g. mass spectrometry experiments on intact proteins. Forthcoming bioimaging experiments with femtosecond X-ray free electron lasers will be aimed at determining the structures of protein molecules \textit{in vacuo},\textsuperscript{38,39} possibly leading to direct experimental determination of protein structures in the gas phase.

\section*{Results}

\subsection*{Constant pressure REMD}

REMD is often performed at constant volume (NVT) but here we have used constant pressure (NPT), in order to avoid high-pressure artifacts at elevated temperatures. The standard Metropolis criterion was used to determine the exchange probabilities:

\begin{equation}
\Delta = \exp(-\Delta U/kT) \leq 1
\end{equation}

\begin{equation}
P(1 \leftrightarrow 2) = \min(1, e^{-\Delta})
\end{equation}

where $\Delta U = U_2 - U_1$ and $\Delta = k_B T \lambda \ln \frac{P_1}{P_2}$ with $T_1$ and $T_2$ the temperatures, $k_B$ is Boltzmann’s constant and $U_1$ and $U_2$ are the potential energies of replicas 1 and 2, respectively.\textsuperscript{40} We have derived a true isobaric–isothermal (NPT) REMD algorithm in which the volume fluctuations are taken into account explicitly:

\begin{equation}
\Delta = (\beta_2 - \beta_1)(U_1 - U_2)
+ (\beta_2 P_2 - \beta_1 P_1)(V_1 - V_2)
\end{equation}

where $P_1$ and $P_2$ are the reference pressures, and $V_1$ and $V_2$ are the volumes in the respective simulations. If we write the components of $\Delta$ (equation (3)) as $\Delta = \Delta_E + \Delta_P$, with $\Delta_U = (\beta_2 - \beta_1)(U_1 - U_2)$ we can consider their relative magnitudes. For the temperature series used here we find that $-2 < \Delta_U < 10$, whereas $\Delta_P \approx 10^{-4}$, since the difference in volume between neighboring replicas is only 0.25 nm$^3$ and the volume fluctuations are of the same magnitude. Hence, we can conclude that the contribution due to the $P \Delta V$ term is negligible, at least as long as the same reference pressure is used for all replicas.

In order to test the effect of the different algorithms on exchange probabilities, a test simulation was performed with constant volume. The REMD temperatures were chosen such as to obtain a constant probability of exchange in accord with equation (2); however, we did not anticipate the density effect on the energy in the NPT simulation. Because of the lower density (Figure 1(d)), at high temperatures the potential energy becomes higher. The energy distributions are given in Figure 1(b) (NPT) and Figure 1(c) (NVT). Due to the reduced overlap of energy distributions in the NPT simulation the exchange probability at the highest temperature is only 1.5\%, considerably lower than the NVT simulation, which has an exchange probability of 8–10\% at all temperatures (Figure 1(a)). Finally, it is interesting to note

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig1}
\caption{Statistics and results from replica exchange MD simulations. (a) Exchange probabilities as a function of temperature. Potential energy distributions for each of the replicas in (b) NPT and (c) NVT simulations. (d) Density in the NPT simulations as a function of temperature, and (e) pressure in the NVT simulations as a function of temperature.}
\end{figure}
pressure as a function of temperature in the NVT simulation (Figure 1(e)). At the highest temperature of 419 K, the pressure is well over 2 kbar.

It is desirable to have similar exchange probabilities over the whole temperature range. What is most important, however, is to have an adequate number of exchanges, high enough such that the algorithm remains efficient but simultaneously low enough to ensure that conformational changes can take place between exchanges. Here (in the NPT production simulations), exchanges were attempted every 2 ps, leading to an exchange rate of 4–22/ ns. Given the slow nature of the folding process this seems sufficient. Application of the true NPT algorithm is preferred, in general, but would not change the results given the temperature range that we selected.

Convergence towards the native state

NMR structures are based on distance restraints derived from the measurement of nuclear Overhauser effects (NOEs). Any distance exceeding the upper limit of the restraint is called an NOE violation (V, measured in nm). Some of the NOEs are due to interactions between groups with more than one proton, e.g. methyl groups, therefore an effective distance d has to be calculated. This was done by summing the distances \( r_{ij} \) between pairs of protons weighted by the sixth power:

\[
d = \left( \sum_i \sum_j r_{ij}^{-6} \right)^{-1/6} \tag{4}
\]

where \( i \) and \( j \) refer to the elements in the individual multi-proton groups. Before computing the violations, the distance were averaged over time. As shown by Tropp, the effective distance between two atoms (or groups of atoms) is not a linear time average. For slowly tumbling molecules (e.g. large proteins) the effective distance is given by:

\[
\tilde{d}_3 = \langle d(t)^3 \rangle^{-1/3} \tag{5}
\]

while for smaller (fast-tumbling) molecules it is:

\[
\tilde{d}_6 = \langle d(t)^6 \rangle^{-1/6} \tag{6}
\]

Since chignolin is somewhere in between a large protein and a small molecule, we have computed both values of \( \tilde{d} \) (Figure 2(a)–(d)). As expected, the time evolution of both quantities is very similar and \( \tilde{d}_6 < \tilde{d}_3 \) always. In all other Figures, we used only the more common \( \tilde{d}_3 \) in order to simplify comparison with other protein simulation results.

The convergence of the REMD simulations to the correct structure was monitored by computing first the distances between atom pairs corresponding to the NOE restraints that were measured (where some atom pairs are taken together according to equation (4)), and, subsequently, the violations, \( \langle V \rangle \). Out of a total of 172 NOE restraints determined for chignolin, 38 were between atoms on residues separated by at least five positions (\(|m-n| \geq 5\)), where \( m \) and \( n \) are amino acid position numbers) and were accordingly classified as long-range restraints, 24 were classified as medium-range restraints (\( 2 \leq |m-n| \leq 4 \)) and 110 were denoted short-range restraints consisting of sequential and intraresidue NOEs (\( 0 \leq |m-n| \leq 1 \)). Long-distance restraints can give rise to large violations (bound only by the length of the polypeptide), whereas the magnitude of short and medium-range restraint violations is limited by the sequence distance. Figure 2(a)–(c) show how nanosecond averages of the long, medium and short-range distance restraint violations converge with time for the simulation of the peptide at 296 K in solvent (note that \( \langle V \rangle \) includes exchanges due to the REMD algorithm). All sets of restraint violations simultaneously converge to low values after approximately 50 ns, indicating that native-like structure has emerged, which is maintained stably after approximately 100 ns. As expected, the magnitude of \( \langle V \rangle \) is highest for the long-range restraints but the pattern of convergence is very similar for all restraint ranges. At lower temperatures, convergence takes longer than at high temperature, but the structures at low temperatures have considerably lower violations,
on average, than the high-temperature ones. The agreement between simulation trajectories and NMR restraints is strong. We find some conformations with average violation (V) less than 0.005 nm, and many with (V) < 0.01 nm. Figure 2(e) shows the distribution of instantaneous (not time-averaged) violations over the last 100 ns of the simulation at 296 K. There is a large peak at (V)=0.01 nm, corresponding to native-like conformations, another smaller hump at (V)=0.06 nm, corresponding to compact structures, and a broad peak around (V)=0.22 nm corresponding to extended conformations. Using (V) as a measure of folding allows a direct comparison of simulation data to experimental measurements. However, Figure 3(a) shows that the instantaneous (V) of all restraints and the more frequently used root-mean-square deviation (RMSD) with respect to the NMR structure (first model from PDB id 1UAO) are almost interchangeable measures of structural similarity.

The REMD simulations in solvent clearly converge from an extended structure to low average (V) values. This “relaxation” process, which is accelerated by the frequent temperature-jumps, takes roughly 50 ns (Figure 2(a)–(d)) for chignolin.

Such a relaxation process has not been observed before in REMD simulations, most likely due to the limited simulation length of earlier studies.

Despite longer sampling (16 µs in vacuo versus 7.36 µs in solvent when adding the length of all 16 trajectories), the native structure never occurs in the vacuum simulations, indicating that the phase space sampled by the peptide is different in solvent and in vacuo. An analysis of the RMSD of each of the trajectories in vacuo (including temperature-jumps) with respect to the final structure shows that the final structure is reached after a few hundred picoseconds, and thereafter only small fluctuations (0.1–0.2 nm Cα RMSD) occur (data not shown). This means that despite the large temperature range, the structure of the peptides collapses rapidly to a certain conformation in vacuo, which is then maintained in the later parts of the simulations presented here.

The individual trajectories of the REMD simulations in solvent reach the native state much more quickly than conventional, single-temperature MD simulations (Figure 3). Due to the temperature exchanges, each replica undergoes many folding and unfolding events, whereas transitions to the native state occur only rarely in conventional MD trajectories. For instance, in two 1.6 µs and 1.8 µs long simulations at constant temperature (300 K) only two to three folding events could be observed in one of the trajectories (Figure 3(c)). The first folded structures were observed after 450 ns, nine times slower than in the REMD trajectories. After 1.2 µs, the peptide unfolds again, and does not refold during the remaining 400 ns of the simulation. The second simulation did not lead to a stably folded structure at all in 1.8 µs (Figure 3(d)).

**Discrete folding events**

If we consider that there are two to three folding events in the first simulation (Figure 3(c)) and none in the second simulation, we have a total of three events in 3.4 µs, which indicates that the folding time is of the order of 1–2 µs. Although one could argue that there is a history in the first simulation, which makes the peptide refold twice, after the initial folding event at 450 ns, the violations in the intermediate unfolded periods (590–650 ns and 1020–1080 ns) are much higher than the smallest violations in the second simulation. Hence, it seems logical that the peptide in the second simulation also should be able to fold completely.

To understand what is going on during these folding events, three indicators are used: the head-to-tail Cα–Cα distance, which measures the degree to which a compact (“globular”) fold has emerged, the distance between the aromatic side-chains (Cg–Cγ), which is indicative of hydrophobic association and the number of main-chain hydrogen bonds, which correlates with the emergence of secondary structure elements. These three parameters thus measure three key properties of natively folded proteins. The results are plotted.

![Figure 3](image-url)
in Figure 4. In the events at 451 ns and 651.5 ns, very similar changes in the observables are found. The number of hydrogen bonds is reduced from two to zero, shortly before the number increases again (to four). The head–tail distance drops in concert with re-formation of hydrogen bonds, while the Tyr-Trp distance decreases only a bit afterward. Hence, it appears that the backbone hydrogen bonds form before the side-chain contacts do. The third folding event at 1083 ns is different, in the sense that the two distances are roughly correct before folding (albeit with relatively high \( V \)), while the number of hydrogen bonds is one higher due to a salt-bridge between the termini of the peptide. Although \( V \) is reduced during this event, neither of the two distances changes significantly, leading to the conclusion this is not a “real” folding event but rather a minor backbone rearrangement involving the termini of the peptide.

**Energy landscapes**

The NOE violations are used to verify the similarity between the simulated and experimentally determined structures. In order to obtain predictive simulations, we need to be able to identify the native state without reference to experimental data. Anfinsen’s dogma, originally introduced as the thermodynamic hypothesis of protein folding states: “…the particular conformation that a protein assumes, under any set of specific conditions, is the one that is thermodynamically the most stable”. 37 Therefore, we compute three-dimensional Gibbs free energy landscapes at a temperature of 296 K and locate the global minima in solvent and in vacuo. In order to discretize the configurational space, two intramolecular distances were used: the head-to-tail distance and the distance between the side-chain atoms \( C' \) of Tyr2 and Trp9. Each of the distances was partitioned in 25 bins with a width of 0.1 nm. In addition the number of main-chain hydrogen bonds was used, which ranges from 0 to 6, yielding a three-dimensional histogram of \( 25 \times 25 \times 7 \) bins. The Gibbs energy of a bin with respect to the global minimum is then given by the probability \( P \) of finding a conformation in each bin:

\[
\Delta G(x, y, z) = -k_B T \ln[P(x, y, z)/P(\text{min})]
\]

where \( k_B \) is Boltzmann’s constant, \( T \) is the temperature and \( P(\text{min}) \) is the probability of finding a conformation in the global energy minimum, which hence has \( \Delta G = 0 \).

For the purpose of this analysis, only the last 100 ns of the simulations were used in order to obtain an equilibrium energy landscape. The Gibbs free energy landscapes for chignolin are plotted in Figure 5. The vacuum landscape is a subset of the solvent landscape, with almost no extended conformations. The global energy minimum in vacuo is located in a different position than in solvent, with larger head-to-tail distance, but comparable Tyr-Trp distance. In vacuo, the global minimum has three backbone hydrogen bonds, whereas there are four in solution. The ensemble at the global minimum consists of 8646 structures which have \( \langle V \rangle = 0.07 \text{ nm} \), an average all-atom RMSD with respect to the NMR ensemble (PDB id 1UAO, 18 structure) of 0.46 nm (min, 0.36 nm; max, 0.51 nm) and an average \( C^\alpha \) RMSD of 0.36 nm (min, 0.29 nm; max, 0.43 nm).

Figure 6 displays the correspondence between Gibbs free energy and NOE violations in solvent at 296 K. The global energy minimum in solvent corresponds to an ensemble of 11,908 structures with \( \langle V \rangle = 0.009 \text{ nm} \), an average all-atom RMSD of 0.19 nm (min, 0.11 nm; max, 0.53 nm) and an average \( C^\alpha \) RMSD of 0.10 nm (min, 0.034 nm; max, 0.39 nm).

The energy landscape presented above is a projection of the high-dimensional energy surface onto three structural order parameters. Consequently, only those structures that fall within the narrow parameter limits of the global minimum bin are included in this ensemble, which thus represents the most frequently observed native-like structures. Rarely observed, high-energy structures that exhibit low \( V \) but have slightly different end-to-end or hydrophobic side-chain distances or...
a different number of hydrogen bonds are grouped into bins surrounding the global minimum in Figure 5. Since the free energy gradient towards the global minimum is steep, native-like structures with low $\langle V \rangle$ can be located in bins with widely varying $\Delta G$ (Figure 6). Most importantly however, the energy minimum corresponds to the native structure.

**Structure prediction**

The projection of the high-dimensional energy landscape onto a three-dimensional surface additionally opens the possibility that some rare, unfolded structures that coincidentally have structural coordinates that correspond to the global minimum are included in this ensemble. In order to exclude these outliers, a cluster analysis was performed as described by Daura et al. Briefly, the all-atom RMSD between all pairs of structures in the global minimum ensemble was determined and for each structure the number of other structures with RMSD of 0.2 nm or less was calculated. The structure with the highest number of neighbors was taken as the center of a cluster. The structures of this cluster were thereafter eliminated from the pool of structures. This procedure produced a cluster of 11,643 structures (97.8% of the structures in the global minimum ensemble) which has $\langle V \rangle = 0.0036$ nm. This compares well with the RMS NOE violations of 0.0018 nm reported for the experimental chignolin structural ensemble. The average RMSD between the predicted and NMR ensembles is 0.18 nm (min, 0.11 nm; max, 0.31 nm)

![Figure 5. Three-dimensional Gibbs free energy landscape $\Delta G(x,y,z)$ (kJ/mol) as a function of two intramolecular distances and the number of main-chain hydrogen bonds at 296 K for (a) the solvent and (b) the in vacuo simulations.](image)

![Figure 6. Correspondence between free energy $\Delta G$ and average NOE violations in solvent at 296 K. The average $\langle V \rangle$ per bin of the Gibbs free energy landscape (Figure 5) is shown.](image)
Table 1. $J_{NNH}$-coupling values from experiment and for the predicted ensemble

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Experimental</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyr2</td>
<td>6.9±1.0</td>
<td>7.3±1.5</td>
</tr>
<tr>
<td>Asp3</td>
<td>8.4±1.0</td>
<td>9.1±0.9</td>
</tr>
<tr>
<td>Glu5</td>
<td>5.8±2.0</td>
<td>6.3±1.7</td>
</tr>
<tr>
<td>Thr6</td>
<td>8.7±1.0</td>
<td>7.2±1.5</td>
</tr>
<tr>
<td>Gly7, H2</td>
<td>5.9±1.0</td>
<td>4.0±1.1</td>
</tr>
<tr>
<td>Gly7, H3</td>
<td>8.6±1.0</td>
<td>8.8±1.1</td>
</tr>
<tr>
<td>Thr8</td>
<td>8.9±1.0</td>
<td>8.3±1.5</td>
</tr>
<tr>
<td>Trp9</td>
<td>8.5±1.0</td>
<td>8.8±1.2</td>
</tr>
</tbody>
</table>

Experimental values taken from the BioMagResBank (http://www.bmb.wisc.edu, accession code 5694). Values were calculated using a Karplus relation with parameters from Plaxco et al.

for all atoms and 0.098 nm (min, 0.034 nm; max, 0.25 nm) for the Cα atoms. For comparison, the average RMSD between different structures within NMR ensemble is 0.13 nm (min, 0.06 nm; max, 0.21 nm) for all atoms and 0.089 nm (min, 0.023 nm; max, 0.017 nm). Additionally, the $J_{NNH}$-coupling parameters calculated for the predicted ensemble are in good agreement with the experimentally determined values, as shown in Table 1.

Temperature dependence

A definition of which fraction of the simulated conformations can be considered native-like ($f_N$) was based on the average distance violations ($V$). A maximum of $V=0.03$ nm (corresponding to a minimum in the distribution curve, see Figure 2(e)) was taken to be the limit to be regarded as folded. Figure 7 shows $f_N$ as a function of temperature for different time stretches of the simulation in comparison to the melting curve as determined by various experimental measures. The simulated $f_N$ is still somewhat too low at very low temperatures and too high at high temperatures. This discrepancy might be caused, in part, by the small size of the simulations box, chosen to minimize the computational cost. The box is sufficiently small to allow some interaction between the peptide and its periodic image in fully extended conformations. Electrostatic interaction between the oppositely charged amino and carboxy termini might thus stabilize unfolded, extended conformations. Nevertheless, the correspondence between simulation and experiment is considerably better than results reported for the β-hairpin from protein G, for the Trp-cage and for a three-helix bundle.

Discussion

Chignolin is the smallest β-hairpin known to be stable in solution. Other examples include an analog of a peptide from tendamistat, and the β-hairpin from protein G. A key feature that these peptides share is the specific interaction between aromatic side-chains. In all cases, including chignolin, interactions between aromatic residues yield an energy surface with a distinct minimum in solution. If there is a well-defined native state, it can be detected by NMR experiments and here we have demonstrated that this structure could be determined independently by MD simulations without reference to experimental data. Small peptides such as these are obviously ideal targets for simulations and, for instance, the peptide from protein G has been studied in detail. Following our analysis of discrete folding events (Figure 4) it seems that the aromatic interactions are not crucial for folding initiation.

The behavior of proteins in vacuo is interesting for comparative studies on protein hydration. Modern mass spectrometry methods allow one to study the effect of “adding” one water molecule at a time to a protein. In practice, the amount of water is determined by the conditions of the “drying” process in vacuo. For moderately charged protein ions, it is expected that the structure in vacuo remains close to the solution structure for some time, although this has not been proved rigorously yet. If proteins unfold in vacuo they do not necessarily fold back to their native (solution) state, as has been indicated by MD simulations. Instead, they collapse to a compact state and remain there without significant rearrangements. In this work, we find that the chignolin peptide does not find its native state in vacuo, despite long sampling and elevated temperatures. Indeed, the global minimum of chignolin in vacuo is very different from the experimental structure (Figure 8). The fact that hardly any conformational change takes place in vacuo after the initial collapse is consistent with earlier simulation results and recent experiments on amyloid-forming peptides. It seems to indicate that a significant external force may be necessary to unfold a protein in vacuo, unless the molecule is highly charged.

It has been suggested that the key to solving the protein folding problem lies in the description of...
the unfolded state. Previous comparisons between solvation models have shown that implicit solvent may not be appropriate to describe the unfolded state, and indeed that water is important to catch details of the folding process itself. Here, we have used explicit solvent and the PME algorithm for long-range electrostatic interactions. These interactions are known to be important for peptide and protein stability. Therefore, the unfolded state, as well as the folding process can be assumed to be modeled accurately in this study.

The simulated melting curve for chignolin (Figure 7) is in good agreement with experiment. Melting curves reported before differed more from experiment and this has been attributed to force field quality. Figure 7 indicates, however, that the melting curve starts to converge only after 200 ns, which is one to two orders of magnitude longer than the REMD simulations used to generate melting curves for the Trp cage (5 ns) and a three-helix bundle (12 ns). Since these two proteins are larger than chignolin, it seems likely that the sampling may have been insufficient for the melting curve to converge. Hence, the interpretation that the temperature dependence of force fields has serious deficiencies must be re-examined. The boiling point of TIP4P water is slightly too low at 363 K, and the critical point is too low as well (588 K rather than 647 K), but simulated water usually remains liquid above the boiling point in constant pressure simulations. Nevertheless, the condensed phases of water are described reasonably well by the TIP4P model and therefore it seems to be justified to use non-polarizable models for biomolecular simulation at temperatures corresponding to the liquid density. It is also conceivable that the incorrect melting curves in earlier work were induced by high-pressure effects due to simulating all replicas at the same volume, rather than to force field problems. This problem is circumvented here by using constant pressure at all temperatures.

Garcia & Onuchic have previously studied the folding of a three-helix bundle using 12 ns REMD simulations with 80 replicas, starting from folded, partially folded, and more or less unfolded conformations. While they were able to study the folding pathway in detail, their findings depended critically on the availability of the native structure. Although our system is smaller, we were able to demonstrate here that the conformations reproducibly fold from an extended state to the native state due to our much longer trajectories, which converge systematically to the correct ensemble (Figure 2). This finding, in combination with earlier promising results, indicates that the protein folding problem can, in principle, be solved by brute force MD simulations, as predicted recently by Karplus & McCammon.  

**Methods**

A linear peptide with sequence GYDPETGTWG was constructed using the PyMOL program and solvated with 888 water molecules and two sodium ions. A rhombic dodecahedron box with periodic image distance of 3.4 nm was used, which has only 0.71 times the volume of a cubic box with the same image distance. This box is large enough to prevent interactions between the folded peptide and its periodic image but sufficiently small to enable some interaction in the fully extended state. The energy of the system was minimized with the steepest descent algorithm, and a 200 ps simulation was performed during which the positions of the protein atoms were restrained. This system was subsequently used as a starting conformation for the replica exchange MD simulations. The OPLS force field was used with TIP4P water. The protonation state of the peptide in solvent was the normal one for pH 7, whereas for the in vacuo simulations all side-chains and termini were neutral. Sixteen replicas were used with temperatures of 275, 282, 289, 296, 305, 313, 322, 331, 340, 350, 360, 371, 382, 394, 406, 419 K, respectively. Simulations were 460 ns in solvent and 1 µs in vacuo. To maintain the temperature at the chosen levels and the pressure at 1 bar, Berendsen weak coupling was used, with coupling constants of 0.1 ps for the temperature and 1 ps for the pressure. A twin range cut-off of 0.9/1.4 nm for van der Waals interactions was applied and the smooth particle mesh Ewald algorithm was used for Coulomb interactions, with a switching distance of 0.9 nm. Neighbor lists were utilized and updated every fifth integration step. Constraints were used for bond lengths using the LINCS algorithm for the protein and SETTLE for the water. In addition, two classical single-temperature trajectories of 1.6 µs and 1.8 µs were generated at 300 K. Hydrogen bonds were defined to be established whenever the angle donor-H–acceptor was less than 30° and the distance H–acceptor distance 0.35 nm or less. All simulations and analysis were performed using the GROMACS software. The simulations reported here required a total of 5.6×10 floating point operations (560 PFLOP), corresponding to roughly 7.5 CPU years on current processors (e.g. Intel Pentium 4 3.0 GHz or AMD Opteron 2.0 GHz).
Acknowledgements

The Swedish grid computing facility SWEGRID is acknowledged for generous allocation of computer time. Janos Hajdu and Abraham Szőke are acknowledged for critical reading of the manuscript. M.M.S. was supported, in part, by Adlerbertska Hospitie-stiftelsen and Stiftelsen Sven och Lilly Lawskis fond för naturvetenskaplig forskning.

References


*Edited by M. Levitt*

(Received 19 May 2005; received in revised form 26 August 2005; accepted 6 September 2005)