doi:10.1006/jmbi.2001.4655 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 309, 299–313

JMB

Essential Dynamics of Reversible Peptide Folding: Memory-free Conformational Dynamics Governed by Internal Hydrogen Bonds

Bert L. de Groot[1], Xavier Daura[2], Alan E. Mark[3] and Helmut Grubmüller[1]*

[1]*Max Planck Institute for Biophysical Chemistry Theoretical Molecular Biophysics Group, Am Fassberg 11, 37077, Göttingen Germany*

[2]*Department of Physical Chemistry, ETH Zentrum Universitätstrasse 6, CH 8092 Zürich, Switzerland*

[3]*Department of Biophysical Chemistry, University of Groningen, Nijenborgh 4 9747 AG, Groningen, The Netherlands*

**Corresponding author*

A principal component analysis has been applied on equilibrium simulations of a β-heptapeptide that shows reversible folding in a methanol solution. The analysis shows that the configurational space contains only three dense sub-states. These states of relatively low free energy correspond to the "native" left-handed helix, a partly helical intermediate, and a hairpin-like structure. The collection of unfolded conformations form a relatively diffuse cloud with little substructure. Internal hydrogen-bonding energies were found to correlate well with the degree of folding. The native helical structure folds from the N terminus; the transition from the major folding intermediate to the native helical structure involves the formation of the two most C-terminal backbone hydrogen bonds. A four-state Markov model was found to describe transition frequencies between the conformational states within error limits, indicating that memory-effects are negligible beyond the nanosecond time-scale. The dominant native state fluctuations were found to be very similar to unfolding motions, suggesting that unfolding pathways can be inferred from fluctuations in the native state. The low-dimensional essential subspace, describing 69% of the collective atomic fluctuations, was found to converge at time-scales of the order of one nanosecond at all temperatures investigated, whereas folding/unfolding takes place at significantly longer time-scales, even above the melting temperature.

Keywords: conformational molecular dynamics; Markov model; peptide folding; principal component analysis; structure prediction

Introduction

Recent computational studies have shown that the problem of peptide (and protein) folding is slowly becoming tractable by atomic-level molecular dynamics (MD) simulations using physically realistic models of these (bio)molecules in explicit solvent (Daggett, 2000; Nardiu *et al.*, 2000; Bonvin & van Gunsteren, 2000; Chipot *et al.*, 1999; Pande & Rokhsar, 1999; Daura *et al.*, 1998; Duan & Kollman, 1998). In the case of some small oligopeptides the simulation times are long enough for reversible folding to be observed and, thus, for a thermodynamic equilibrium between the folded and unfolded states to be established. One of the most extensively studied cases so far has been the reversible folding of a non-natural seven residue β-peptide (composed of β-amino acid residues) which folds into a left-handed 3_1 helix in methanol (Daura *et al.*, 1998). MD simulations have been performed at four different temperatures ranging from 298-360 K, and the resulting trajectories, spanning up to 200 ns, correspond well with the available experimental NMR and circular dichroism data (Daura *et al.*, 1997, 1999a,b; Seebach *et al.*, 2000).

These simulations thus provide a unique opportunity to study peptide folding at the atomic level, and can be considered a model system for the study of the folding of larger polypeptides. The folding/unfolding simulations have been characterized in detail from a structural perspective. In particular, the folding pathways of this β-peptide

Abbreviations used: MD, molecular dynamics; PCA, principal component analysis; HBN, hydrogen bond network; MSF, mean-square fluctuation.

E-mail address of the corresponding author: hgrubmu@gwdg.de

0022-2836/01/010299–15 $35.00/0

have been analyzed, the relative free energies of different conformational sub-states have been estimated, and the correlation of several geometrical properties to the degree of folding of the peptide has been investigated. One of the surprising results was that the unfolded state as well as the folding process involve a relatively small number of conformational states (Daura *et al.*, 1998, 1999b).

Extending this analysis, we have investigated in detail the structural and the dynamic properties of the transitions between the different folding (sub)-states. For this purpose, a principal component analysis (PCA) was performed, which yields the main collective fluctuations observed in a cluster of structures. Applied to the peptide folding simulations, it shows how many collective degrees of freedom are necessary to adequately approximate the folding dynamics, depending on the complexity and diversity of the folding transitions.

For a coarse-grained description of the peptide dynamics to be successful, a small subset of relevant collective degrees of freedom must exist that allow a sufficiently accurate description of the folding motions. In addition, the remaining (and neglected) degrees of freedom must be negligible in terms of both structure and dynamics. In particular, their influence on the essential degrees of freedom, possibly showing up as memory effects, should be small. With the relatively long simulations available, these two main prerequisites for the application of such dimension-reduced models for the description of long-time dynamics of macromolecules can now be tested.

Sampling must be dominated by a small number of collective modes

If indeed the majority of the collective fluctuations is found to take place along a small number of collective degrees of freedom, the simulations of the folding/unfolding equilibrium of this peptide with a length of 200 ns allow an investigation of the speed of convergence of these coordinates. The question of whether a principal component analysis is suitable to investigate protein dynamics from simulations in the time-scale of hundreds of picoseconds to nanoseconds has been under debate for some time. Some authors have argued that the low-dimensional "essential" subspace in which the majority of the collective fluctuations are often found to take place would not be converged at such short time-scales (Balsera *et al.*, 1996; Clarage *et al.*, 1995), whereas others found that a useful approximation of the essential subspace can usually be obtained after relatively short simulation times (De Groot *et al.*, 1996b; Amadei *et al.*, 1999a). Moreover, applications to native-state dynamics of proteins have shown that dimension-reduced models derived from PCA can be useful in the interpretation of simulated protein dynamics (García, 1992; Amadei *et al.*, 1993; van Aalten *et al.*, 1995a,b; Hayward *et al.*, 1995) and of dynamical properties derived from clusters of experimental structures (de Groot *et al.*, 1998; van Aalten *et al.*, 1997; Abseher *et al.*, 1998). A PCA of simulations of the unfolding of a β-hairpin has provided an initial indication that dimension-reduced models are also useful in the description of the (un)folding dynamics of peptides (Roccatano *et al.*, 1999). If the full dynamics of a system can be approximated successfully by relatively few global degrees of freedom, this can be exploited in techniques that stimulate enhanced sampling of these coordinates (Amadei *et al.*, 1996; de Groot *et al.*, 1996a; Grubmüller, 1995; Abseher & Nilges, 2000). The long simulations available for this peptide provide a good opportunity to study the convergence and stability of this subspace, and its dependence on the starting conformation in short pieces of trajectory that suffer from incomplete sampling.

Memory effects must be small on the time-scale of the dynamical process of interest

If long-time memory effects play a major role in large-scale conformational transitions, then short-time simulations will not provide insight into long-time behavior of the same system. If, however, such memory effects are small, then coarse-grained descriptions of the folding dynamics derived from short simulations may be valid. To decide if this is the case for the peptide at hand, two independent tests for memory effects in conformational transitions are presented. The first involves a comparison with a model that assumes a Markov process for transitions between stable structures based on a clustering in the configurational space, the second is more general, in that it is independent of any particular choice of clustering.

If the above two conditions are fulfilled for the simulations of the β-peptide, then the distribution of conformations in the essential subspace should structurally and dynamically reflect and characterise concepts like folding intermediates and folding pathways. Moreover, given an acceptable level of convergence, the free-energy landscape can be mapped directly onto the principal coordinates, as the relative configurational free energies are related directly to the density of states in this subspace. Such an analysis would then indicate conformational sub-states along the "natural" folding coordinates, with barriers of varying height in between. Furthermore, the entropic contribution could be estimated from the difference in behavior of the simulated system at different temperatures.

Geometrical properties that correlate with the degree of folding

An interesting aspect of folding in general, and for native structure prediction in particular, is the search for geometrical properties that correlate with the degree of folding. Daura *et al.* (1998) have shown that all of the geometric or energetic properties investigated correlate poorly with the root-mean-square deviation (RMSD) from the

experimental structure of the β-peptide. Among the properties investigated, the number of 3_1 backbone hydrogen bonds was found to be most informative, in the sense that structures with five such hydrogen bonds all had a native conformation. However, structures were found with zero or only one such hydrogen bond that structurally resemble the native conformation. Therefore, the number of native backbone hydrogen bonds alone was not diagnostic of the folded state. Here, we investigate if a knowledge-based hydrogen bond network potential improves the accuracy of identifying the

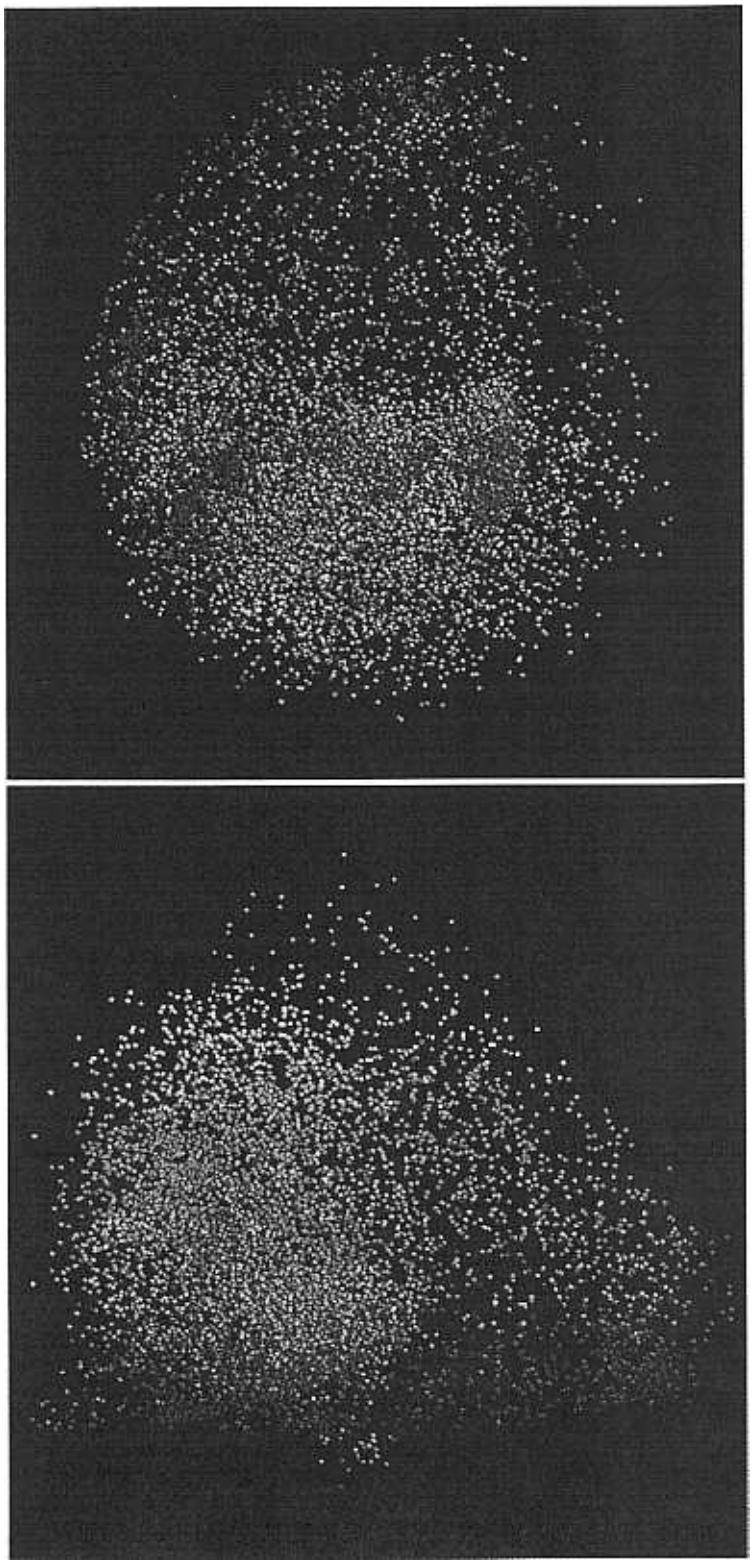

Figure 1. Projections of the 340 K simulation onto two planes spanned by a linear combination of the three eigenvectors with largest eigenvalues of an all-atom PCA, coloured according to the radius of gyration of the peptide (blue for small gyration radius, red for large). Despite the clear correlation between the projection onto these coordinates and the radius of gyration, there are many, structurally diverse, conformations with relatively small gyration radii.

native conformation among a cluster of folded, partially folded and unfolded conformations.

Results

Structures

Figure 1 shows two different projections of the simulation at 340 K onto two suitably chosen planes defined by linear combinations of the three eigenvectors with largest eigenvalues from an all-atom PCA. Thus, each "ball" in the picture represents one snapshot from the simulation. The colour denotes the radius of gyration of each snapshot. The extreme values of the gyration radius are found near the borders of the space spanned by these degrees of freedom. On one side are the completely stretched conformations (in red), sampling a relatively small fraction of configuration space, whereas on the other side there are many structurally diverse conformations with similar, low gyration radii (in blue). Overall, the sampled subspace in these degrees of freedom resembles a half-sphere, with the top corresponding to the most stretched conformations, and the flat bottom corresponding to all compact conformations, of which the native left-handed helix forms a small, dense fraction.

Figure 2 shows the projection of the 340 K simulation onto the plane defined by the two principal eigenvectors of the covariance matrix constructed from the central backbone coordinates. Here, the projections are coloured according to the configurational density, calculated from nearest-neighbor distances as described in Materials and Methods. The largest, most densely populated sub-cluster (I) corresponds to the native left-handed helix (Figure 3,I), with all helical backbone hydrogen bonds intact between residues i and $i+2$. The closely connected, but distinct dense sub-cluster (II) corresponds to a conformation similar to the native structure, but lacking the two C-terminal backbone hydrogen bonds (Figure 3,II). The third-most dense sub-cluster (III) corresponds to a hairpin-like structure (Figure 3,III) in which the C terminus folds back onto itself, forming hydrogen bonds between the backbone of residues 2 and 6 as well as between 3 and 6. The N terminus in this structure is folded back onto the turn, and the terminal amino group forms hydrogen bonds with residues 2, 4 and 5. The remaining, relatively diffuse, cloud of structures consists of unfolded and partially folded conformations. The three dense clusters are also identified by k-medoid clustering applied to the 60 coordinates of the central backbone (blue, cyan, yellow, respectively, in Figure 4(a); the remaining conformations are depicted in red). An almost identical definition of the clusters is obtained when the clustering is limited to the three-dimensional subspace spanned by the three

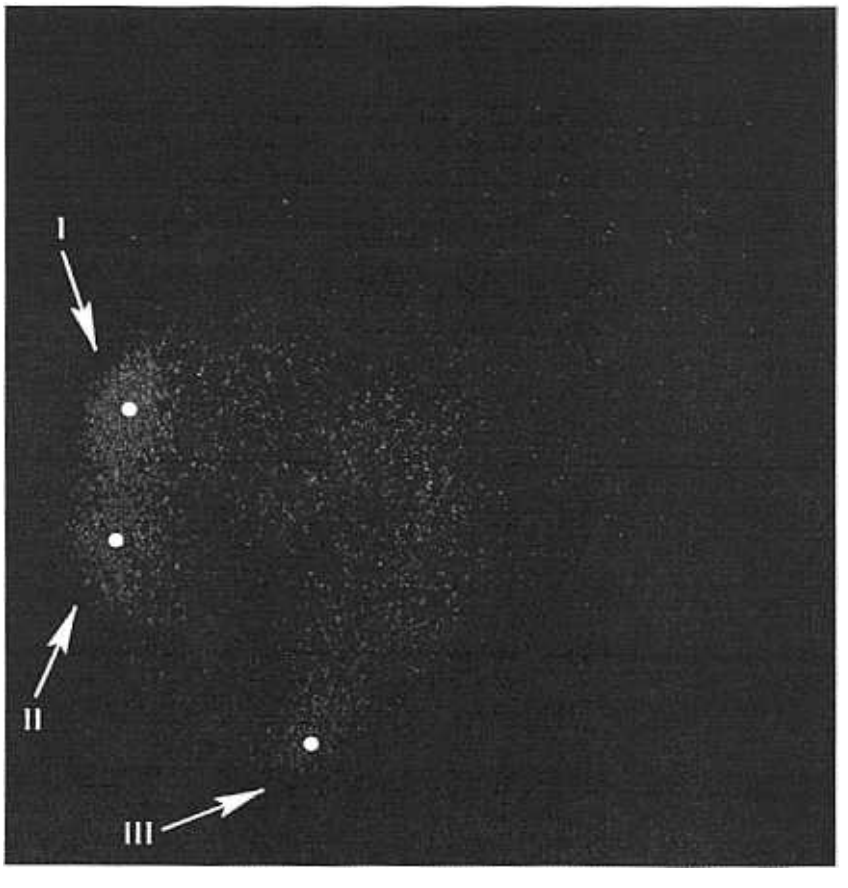

Figure 2. Projection of the 340 K simulation onto the plane spanned by the two principal eigenvectors of a PCA performed over the central backbone coordinates, coloured to configurational density. Indicated are the three regions of highest density. The central structures of each of these three densest clusters (indicated by yellow circles) are depicted in Figure 3. Apart from these three dense sub-clusters, the cloud of (unfolded) conformations is relatively unstructured.

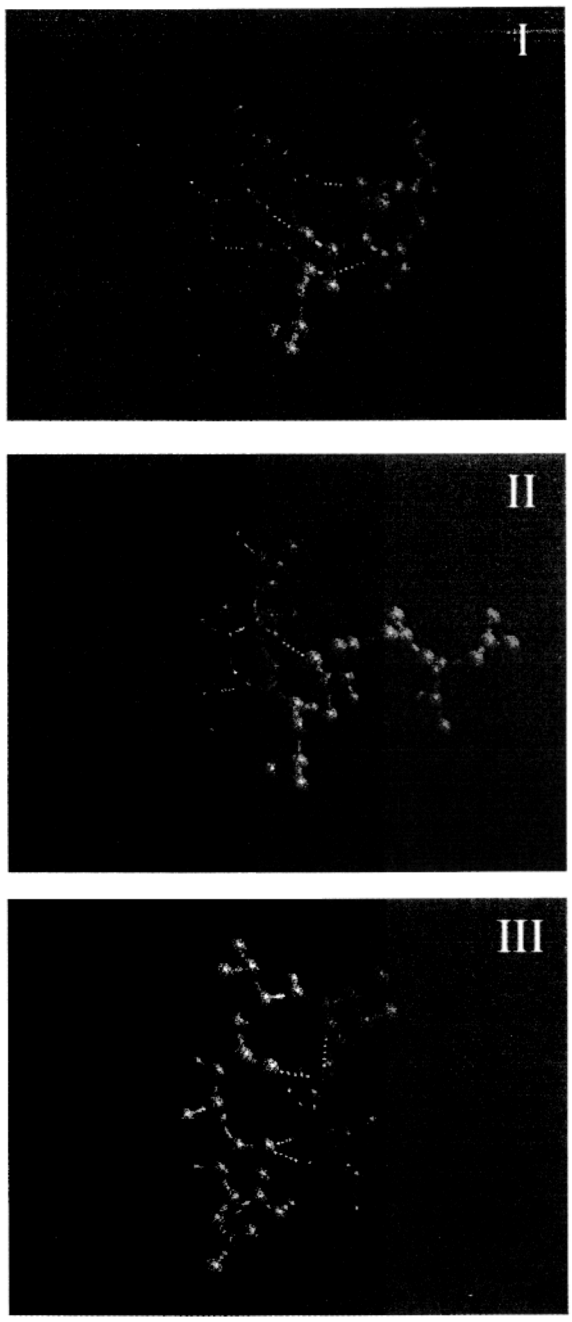

Figure 3. Structures of central cluster members of the three densest clusters (see Figure 2). The structures are coloured blue to red from the N to the C terminus. Indicated by broken bars are internal hydrogen bonds. This Figure was made with Bobscript (Esnouf, 1997; Kraulis, 1991) and Raster3D (Merritt & Bacon, 1997).

eigenvectors of the covariance matrix with the largest eigenvalues, which account for 69% of the total fluctuation in the simulation (Figure 4(b)).

Dynamics

Transitions between the four clusters as observed in the 340 K simulation were monitored and a corresponding transition frequency matrix was constructed. From the transition frequencies, the optimal Markov rates were calculated as described in Materials and Methods. The Markov model obtained is depicted in Figure 5. The ratios between the transition probabilities correlate well with the corresponding occupancies of the involved states. Only for the I → III transition the transition probability might be underestimated, since no transition of that kind has been observed, although a value of 0.1 would be expected from the III → I transition frequency and the occupancy difference of both states. Several "simulations" of the true Markov process based on these parameters were compared to the MD simulation. In order to validate whether the folding/unfolding transitions in the MD trajctories can be described by the (memory-free) Markov model, conditional transition probabilities were monitored as a function of time both for the MD simulation and for the simulated Markov trajectories. Except for the III → III transition probability (the probability to be in state III at time t and at time $t + \Delta t$), all the 16 first-order conditional transition probabilities are, within statistical error, identical for the MD simulation and the simulated Markov processes (Figure 6). This implies that, overall, the folding transitions between the major folding states can be described successfully by a Markov process. Memory effects appear to play only a minor role during the folding of this peptide.

An analysis of the transition frequencies between the clusters shows that 95% of the transitions towards cluster I, the native left-handed helix, originate from cluster II. Together with the structural similarity of the two states, this suggests that cluster II can be described as a folding intermediate. Just over 3.5% of the transitions towards cluster I originate from cluster IV, the collection of unfolded conformations and approximately 1.5% from the hairpin-like structure of cluster III.

As described in Materials and Methods, memory effects were also investigated independently of any particular clustering. Figure 7 shows the distribution of times for transitions that return to a state that had been sampled immediately before (21,722 transitions), compared to the distribution of all conformational transitions (without the requirement of returning to a previously sampled configuration, 1,169,873 transitions). For comparison, both distributions were normalised to 1. The largest difference is observed for transitions taking place at time-scales of less than 1 ns, indicating a memory effect that causes the peptide to return to conformations that have been sampled shortly before. The apparent differences observed at time-scales from 4-8 ns are believed to be caused by the dominance of the short-time memory effect in the normalization procedure.

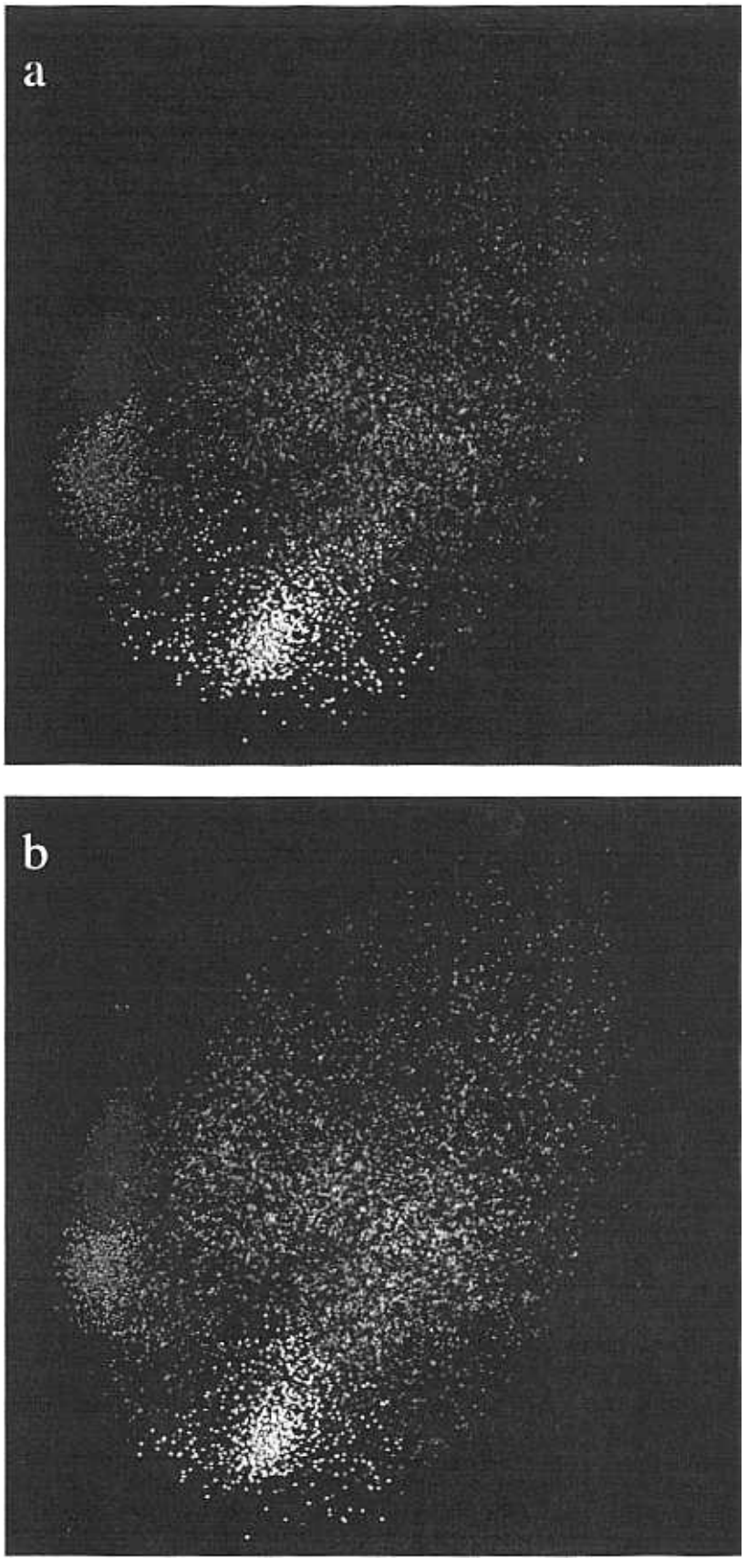

Figure 4. Results of a k-medoid clustering (a) in the full 60-dimensional coordinate space spanned by the the central backbone residues and (b) in the subspace spanned by the three principal eigenvectors calculated for these residues.

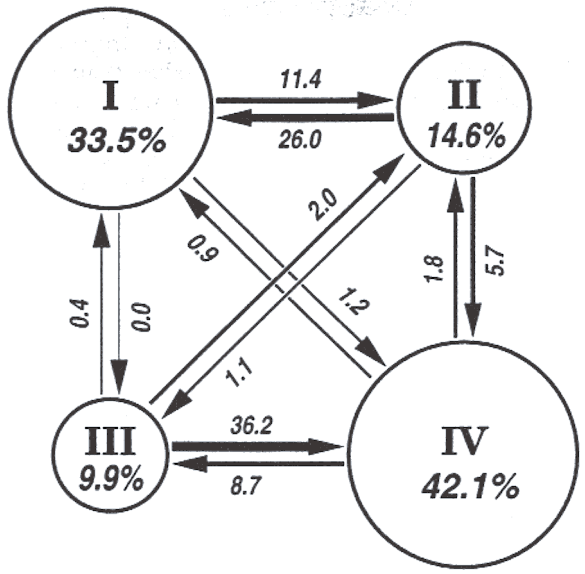

Figure 5. The four-state Markov model that best describes the transitions between the core parts of the four clusters derived from the 340 K simulation. The numbers along the arrows denote transition probabilities ($\times 10^4$) between the clusters in time-steps of 0.5 ps, the numbers inside the circles denote the occupancy of each cluster.

Free energies

Relative configurational free energies were calculated as the logarithm of the configurational density in the 60-dimensional space spanned by the central part of the backbone residues (see also Materials and Methods). Figure 8 shows the distribution of the calculated free energies from the 340 K simulation along the collective backbone fluctuation with the largest amplitude (eigenvector of the covariance matrix with the largest eigenvalue). Therefore, the three high-density clusters that had been identified by the clustering procedure are recognizable as the lowest free-energy conformations. Clusters I (the native left-handed helix conformation) and II (the main folding intermediate) are related structurally and have similar free energies, with only a relatively small barrier between the two states. There are no low free-energy pathways connecting cluster III to clusters I and II, as deduced from the transition rates between these clusters. For successful folding to take place from cluster III, the structure will first have to undergo a conformational change towards a less structured state of higher free energy, before it can re-fold towards the native structure.

Figure 9 shows a comparison of the relative sampled configurational free energies for the simulations at 298 and 340 K. The simulation at 298 K samples clusters I and II more frequently than the simulation at 340 K, leaving the unfolded state poorly sampled; cluster III is not visited at all. In the simulation at 298 K, the difference between clusters I and II is more pronounced than at 340 K, with the barrier between the two clusters being clearly visible. This suggests an entropic contribution to the free-energy difference between the two most densely sampled clusters.

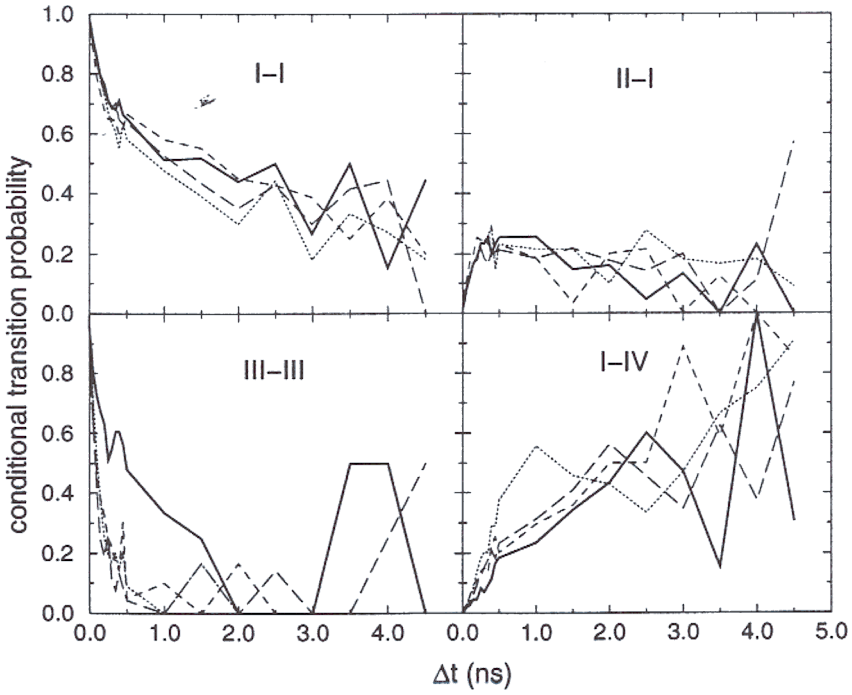

Figure 6. Conditional transition probabilities between clusters in the 340 K MD simulation (continuous, bold line) and in simulated Markov processes (broken lines) as a function of time. The simulated Markov processes were based on the transition matrix depicted in Figure 5 and differ from each other in the applied random-number seed. From the scatter of the three Markov probabilities, the statistical error can be estimated.

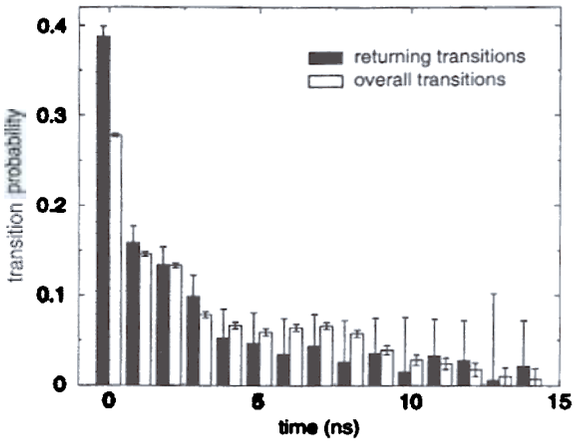

Figure 7. The distribution of transition times for conformational transitions in which the system returns to a state that had been sampled immediately before (filled bars), compared to the distribution of transition times of similar conformational changes, but now without the condition to a previously visited state (open bars). The error bars correspond to the statistical standard deviation estimated from the observed transition frequencies.

Folding determinants

The observation that the major conformations involved in the folding of this peptide have markedly different hydrogen-bonding patterns led us to investigate the internal hydrogen bond network (HBN) energies in detail. Figure 10 shows plots of the HBN energies of conformations extracted from the simulation at 340 K against the RMSD from the experimental structure and against the relative configurational free energies. In line with the results from a previous analysis, which compared the number of native backbone hydrogen bonds to the RMSD from the "native" structure (Daura *et al.*, 1998) rather than energies, both graphs show a clear correlation between the HBN energies and the distance from the native structure (with correlation coefficients of 0.72 and 0.76, respectively).

For any geometrical or energetical property to be useful as a predictive measure for the degree of folding of any peptide or protein, the correlation between the calculated property and the degree of folding should be high, and the number of outliers (false positives and false negatives) should be small. To investigate how reliably the HBN energies can identify the native conformation in the ensemble of structures simulated at 340 K, the number of native and non-native conformations with HBN energies above and below a certain energy threshold were counted for different threshold values. As can be seen in Figure 11, for values of the HBN energy of around 3.0, the probability of correctly identifying the native state based on HBN energies is about 85%, since the probability that conformations with an HBN energy higher than 3.0 are native (or positive predictive value) and the probability that conformations with an HBN energy smaller than 3.0 are non-native (negative predictive value) are both approximately 86%. For higher values of the HBN energy threshold, the probability of correctly identifying the native state is even higher, but now only for conformations with an HBN energy larger than the threshold.

Convergence of essential subspaces

Figure 12(a) shows that for the 340 K simulation, the three-dimensional essential subspace has reached an almost complete convergence after 200 ns. The overlap between the essential subspaces extracted from the first and last 90 ns with the full 200 ns trajectory is 98.6% and 98.3%, respectively. The overlap between the first and last 90 ns is 94.3%. Moreover, the subspace converges relatively fast: already within 100 ps, an overlap with the converged set of eigenvectors of 58% is reached on average, and within 10 ns an overlap of 80% is accomplished. In terms of atomic displacements, the three-dimensional essential subspace extracted from the central backbone configurations of the 340 K trajectory accounts for 69% of the total (internal) mean-square fluctuation. Figure 12(b) shows that essential subspaces extracted from 100 ps trajectories, on average, account for 63% of these collective fluctuations, making up 43% of the total mean-square fluctuation. For trajectory pieces of 10 ns, these numbers are 88% and 61%, respectively.

To investigate how similar the unfolding motions are to the native state fluctuations, a PCA was performed over all structures of cluster I (see Figure 4) and subspace overlap values were calculated with respect to the converged set of eigenvectors extracted from the 200 ns simulation. The essential subspaces show an overlap of 52% (Figure 13) and the subspace spanned by the first six eigenvectors of the native state cluster reproduce the essential subspace of the complete trajectory for 82%. From the sixth eigenvector on, the convergence to 1 occurs at a similar rate as is observed for 1-ns fragments of the full trajectory. This indicates that unfolding directions are highly similar to native state fluctuations. The largest contribution to the three main unfolding directions comes from the native state essential subspace. The next three degrees of freedom in terms of the amount of fluctuations in the native state form the second largest contribution to the overlap with the principal unfolding directions. This indicates that unfolding takes place mainly along degrees of freedom that already exhibit an appreciable amount of fluctuation within the native state cluster, and hardly involves directions that are not accessible in the native state.

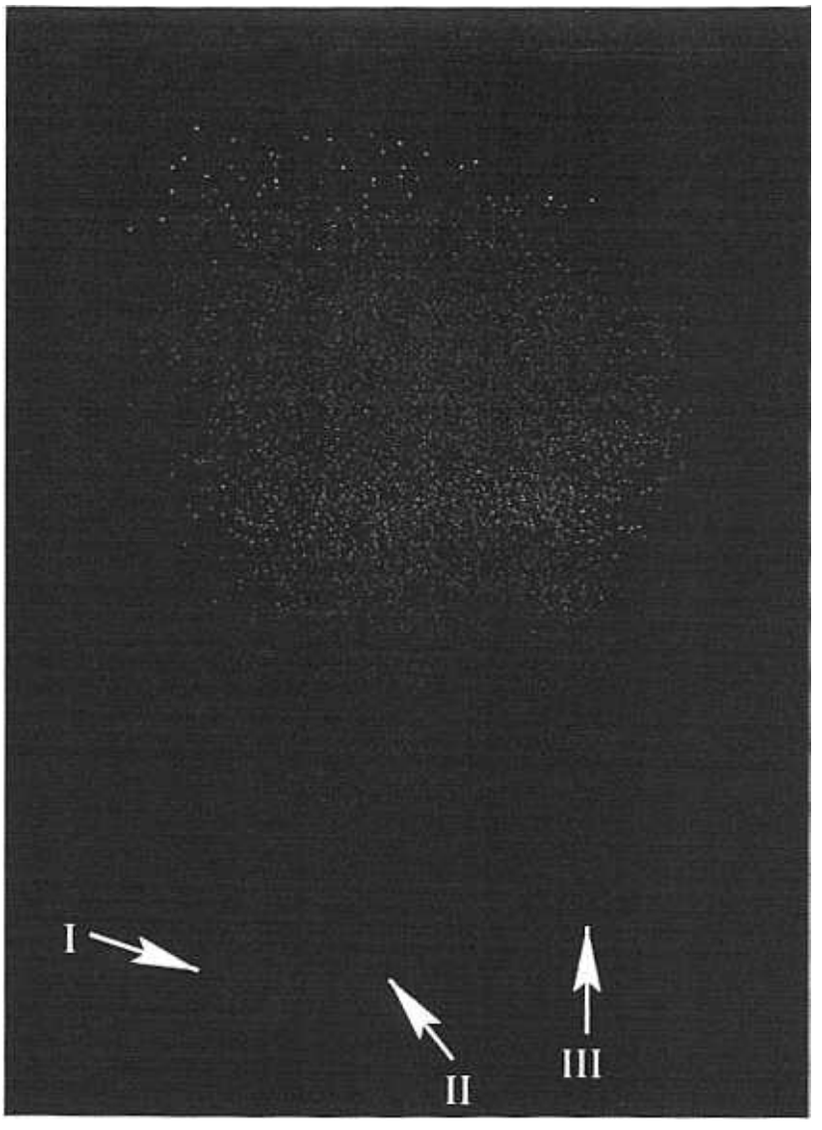

Figure 8. Relative configurational free energy distribution of conformations extracted from the 340 K simulation along the principal coordinate derived from the central backbone atoms. Relative free energies were calculated from the phase-space densities, derived from averaged nearest-neighbor distances (RMSDs calculated over the central backbone atoms). The vertical coordinate as well as the colour indicate the free-energy value (with the lowest free-energy values blue). The three dense sub-clusters are indicated in yellow. The *x*-axis is a linear combination of the two principal eigenvectors that span the plane shown in Figure 2, and was chosen to optimally show the difference between the clusters.

Discussion

The PCA results of the equilibrium folding dynamics of a β-peptide presented here provides a perspective of the conformational dynamics that highlights different aspects from that of the RMSD-based analysis (Daura *et al.*, 1998, 1999b). The borders of the essential subspace correlate well with extrema in the gyration radius of the peptide (cf. Figure 1), indicating that global boundaries as defined by the MD force-field correlate with the degree of compactness of the peptide. On one side, the most stretched conformations are found, indicating that a physical border of the force-field connected to the stretching of covalent bond lengths and angles is approached. On the other, flat side of the half-sphere, the most compact structures are sampled, of which also the native state cluster is part. The relatively sharp border of the essential subspace on this side corresponds to sharp repulsive Lennard-Jones energy terms that prevent substantial non-bonded atomic overlap beyond the van der Waals radii.

Apart from the three dense sub-clusters (Figures 2 and 4), and in contrast to the larger number of clusters obtained by Daura *et al.* (1999b), the essential subspace appears to be sampled relatively homogeneously, which raises the question of whether a further subdivision in new clusters is meaningful. Moreover, the comparison of the clustering carried out in the space spanned by the three principal eigenvectors and in the full configurational space shows that virtually all features are captured already in this low-dimensional essential subspace. Therefore, structural differences in the orthogonal subspace play only a minor role and, for most purposes, a projection onto only the first principal modes yields a satisfactory approximation of the dynamics in the full configurational space.

One feature of the ensemble revealed in the PCA results is an intermediate state (state II in Figures 2

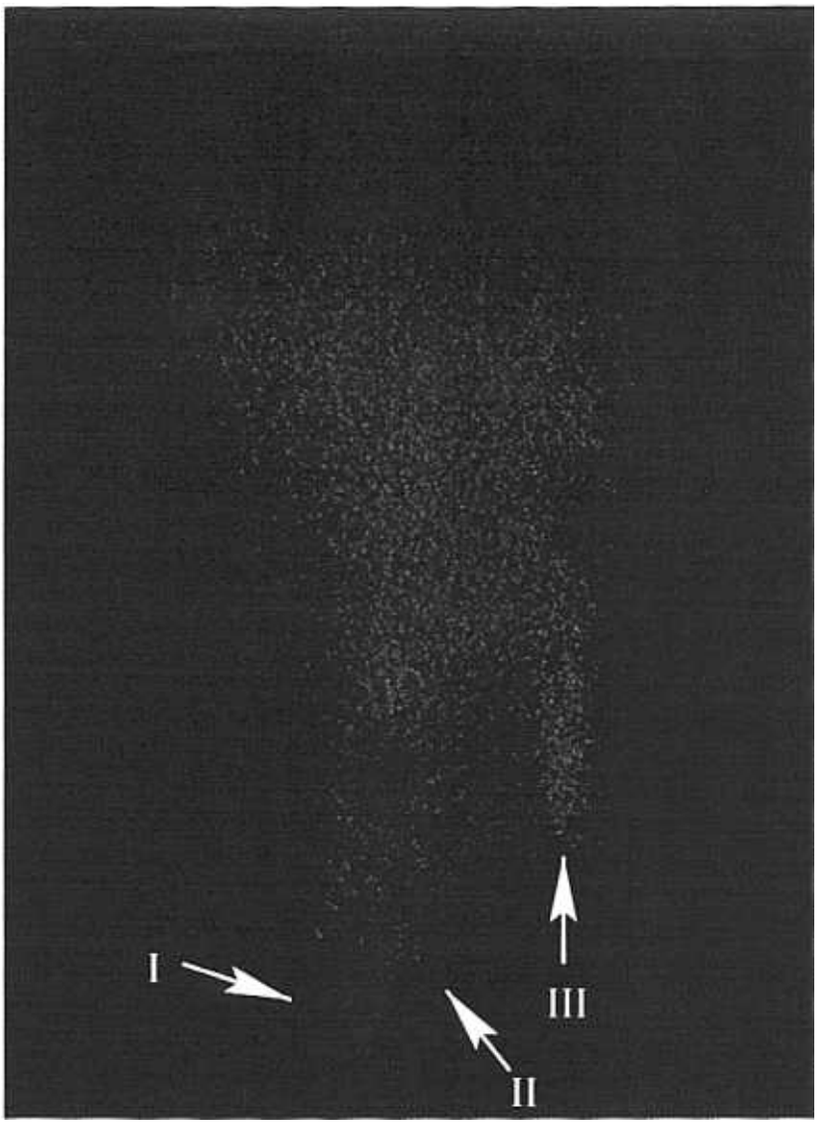

Figure 9. Relative configurational free energies (vertical axis) at different temperatures along the principal coordinate of the 340 K simulation (horizontal axis). The simulation at 298 K is depicted in blue, the 340 K simulation is depicted in red. The 298 K simulation clearly shows the free-energy barrier between clusters I and II, whereas the 340 K simulation shows that there is no low free-energy pathway connecting cluster III to the native state (see also Figure 8).

and 4) close to the native state. Because of the remarkably small RMSD between states I and II of only 0.8 Å this intermediate was not seen in the 1 Å RMSD clustering described before (Daura *et al.*, 1999b). Nevertheless, the barrier between the two states is clearly visible in Figure 9 and indicates that indeed there are two structurally related but distinct conformational states.

Both the k-medoid clustering presented here as well as the RMSD clustering suffer from a general problem of clustering algorithms: eventually all data points will be clustered, even if the inherent structure of the distribution may not justify it. In some cases, this can cause an overestimation of the amount of substructure in a dataset that is hard to recognise. The k-medoid clustering as presented here also suffers from this drawback, and the significance of the clusters can be estimated only from the dependence of the cluster definition on the chosen parameters, most notably the number of clusters to be defined. More sophisticated clustering algorithms like fuzzy clustering may overcome this problem partially, but we feel that an optimal clustering for datasets of a dynamical origin should take into account the static distribution of data points, and at the same time optimize residence times and transition frequencies between clusters.

The only significant deviation in the simulation of 340 K from a Markov process shows up in the conditional probability to find the system in state III (Figure 6). For the first 2 ns, the probability to remain in (or return to) cluster III is higher than expected from the corresponding Markov process. A possible explanation for this effect is that clusters III and IV are relatively close to each other, and the barrier between them is small. Therefore, the system can easily cross the border from III to IV and return without having been unfolded completely, and thus without actually having reached cluster IV. In the simulated Markov processes, all transitions are sharp and complete. Hence, with the same transition probabilities, the chance to return

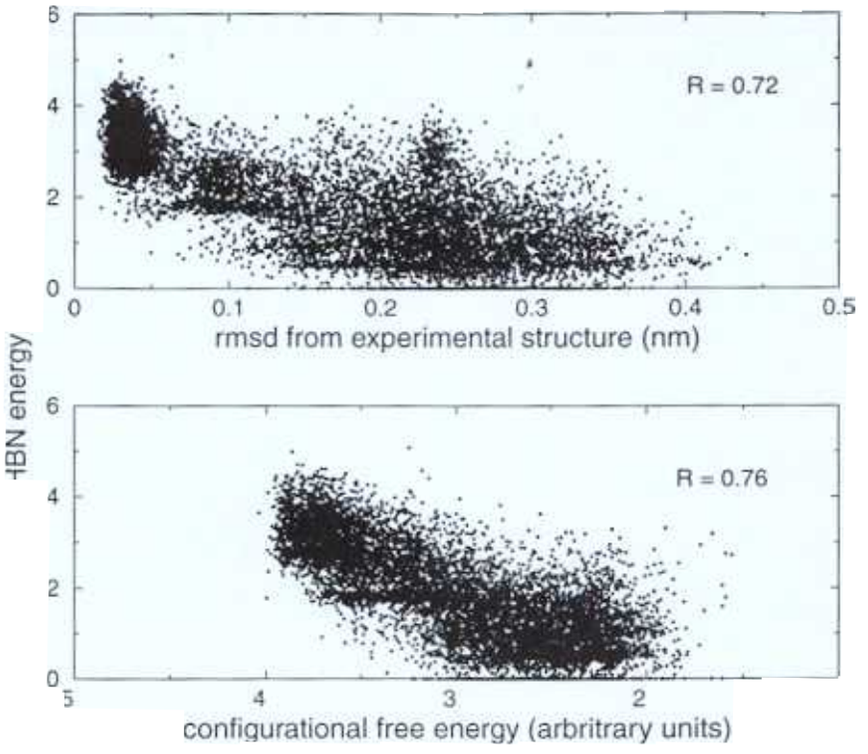

Figure 10. HBN energies calculated from snapshots from the simulation at 340 K compared to the backbone RMSD from the experimental structure (upper panel) and to the configurational free energy (lower panel). Indicated are the correlation coefficients between the two datasets.

to cluster III is smaller in the simulated Markov processes than in the MD simulation.

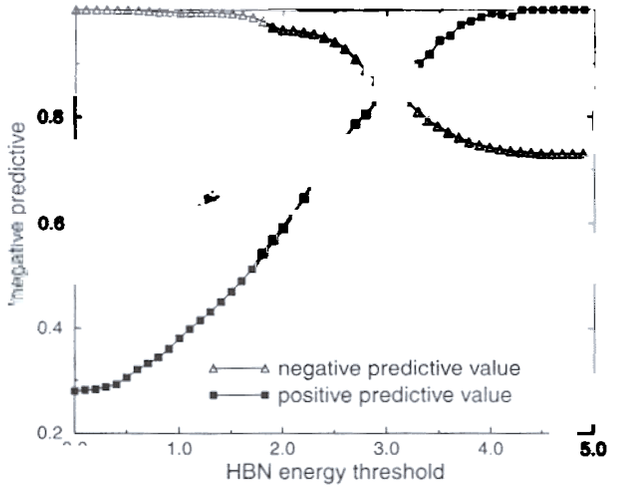

Figure 11. Positive and negative predictive values (probabilities of correctly identifying native and non-native structures, respectively) based on HBN energies at different energy threshold values. The curve with the filled squares depicts the native-state fraction of all conformations with a HBN energy above the threshold. The curve with the open triangles depicts the non-native fraction of all conformations with a HBN energy below the threshold. For HBN energies of around 3.0, the probability of successfully identifying the native as well as the non-native state is about 85%.

The PCA results have provided an intuitive basis for the presentation of calculated conformational free energies (Figures 2 and 9), clearly showing the relative weights of each cluster and the barrier heights between the clusters. The kinetics, deduced from the Markov process based on the transition rates between the clusters, agree perfectly with the picture obtained from the PCA.

Although some outliers are found when the HBN energies are plotted against the conformational free energies or against the RMSD from the experimental structure (Figure 10), the HBN energies are a more sensitive and unbiased measure of the folding state of this peptide than the number of native backbone hydrogen bonds. Moreover, the obvious correlation between the HBN energy and the configurational free energy suggests that internal hydrogen bonding plays an important, and possibly dominant, role in stabilizing specific conformations and guiding the folding of this peptide.

We have found the essential subspace to converge fast, as can be seen from the average subspace overlap of 0.58 from fragments of only 100 ps with the full 200 ns trajectory at 340 K. Of the mean square fluctuation that takes place in the essential subspace extracted from the 200 ns simulation, 63% can be described by essential subspaces from trajectory fragments of 100 ps, on average. Such fast convergence may not be transferable directly to the dynamics of more complex proteins,

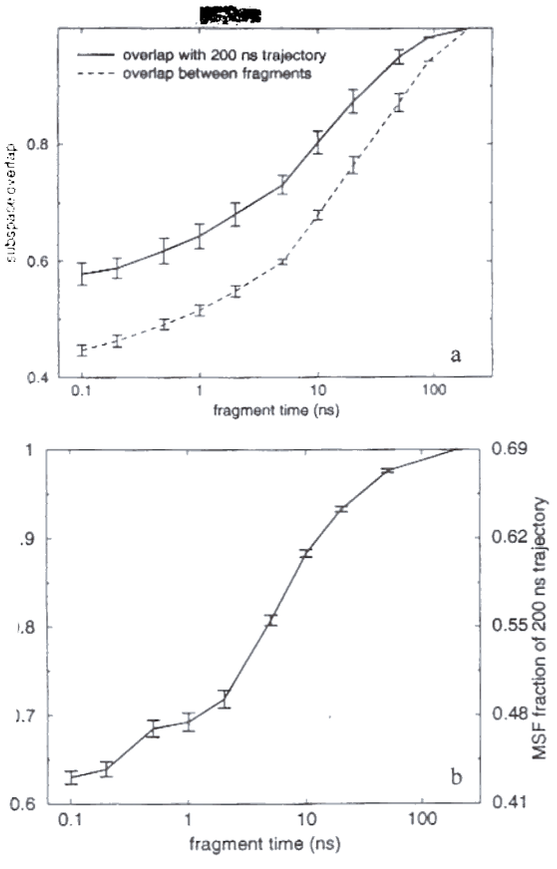

Figure 12. Convergence of the essential subspace of the 340 K simulation. (a) Essential subspace overlap values (see Materials and Methods for a definition) between fragments of varying length with the converged set extracted from the 200 ns simulation (continuous, bold line) and between pairs of fragments mutually (broken line). (b) Fraction of the mean-square fluctuation (MSF) of the 200 ns trajectory at 340 K that can be described by a three-dimensional essential subspace extracted from trajectory fragments of different length. Both the overall fraction is shown (right-hand side) and the fraction of the fluctuation described by the three-dimensional essential subspace extracted from the 200 ns trajectory (which describes 69% of the total MSF).

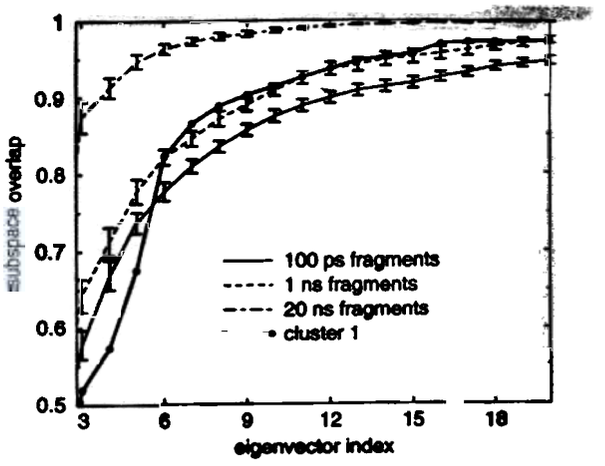

Figure 13. Comparison of native state fluctuations to unfolding motions. The bold, continuous line depicts the subspace overlap between eigenvectors extracted from the native state cluster to the overall 200 ns simulation at 340 K. For comparison, the same type of graph is shown for fragments of varying length from the full trajectory.

either in their native state or during folding. However, the facts that the essential subspace converges significantly faster than the typical folding time, and that the essential subspace of one sub-state yields a reasonable approximation of the overall subspace, indicate that dimension-reduced models will prove valuable in the study of (folding) dynamics of macromolecules.

Two similar models have been proposed to describe conformational transitions in proteins at different time-scales. In both, the short-time behavior is described by motion within a harmonic well. In the first, the long-time behavior is described by a hopping among the harmonic minima (Kitao *et al.*, 1998), whereas in the second the transitions between the minima are of a diffusive nature (Amadei *et al.*, 1999b). Central in both models is the similarity between principal directions within one minimum to conformational transitions between minima. The fact that the essential subspace derived from the native sub-state shows a significant overlap (82%) with the overall folding/unfolding essential subspace suggests that the concepts used in both models apply also to folding transitions between conformational states that are themselves collections of local minima.

Conclusions

The results presented here show that dimension-reduced models can be used to describe many aspects of conformational transitions involved in peptide folding. It has been demonstrated that the two main prerequisites for such an approximation, a fast convergence of the essential subspace and negligible memory effects at large time-scales, are fulfilled. For the β-peptide investigated, the set of unfolded conformations comprises a relatively diffuse cluster of structures in configurational space flanked by three more dense clusters identified as the native fold, the main on-pathway folding intermediate and a non-productive off-pathway conformation. Therefore, the number of sampled conformations during folding is much smaller than could be expected (Daura *et al.*, 1998), and native-state dynamics as well as folding dynamics involves primarily only a few of the many apparently available collective degrees of freedom. This should simplify the search for the lowest free energy state considerably. In addition, the finding that internal HBN energies correlate significantly

with the degree of folding opens new possibilities for the design of simplified interaction potentials targeted towards peptide or protein folding.

Materials and Methods

Molecular dynamics simulations of a β-heptapeptide in solution at different temperatures were analyzed. Details of the simulations have been described (Daura *et al.*, 1998). In total, four simulations were analyzed: at 298 and 340 K of 200 ns each, and at 350 and 360 K, of 50 ns each.

Principal component analysis was carried out by diagonalizing the positional covariance matrices constructed from the trajectories after a least-squares fit to a reference structure. A PCA was performed on all atoms of the peptide and on the central backbone (residues 2-6) using the WHAT IF (Vriend, 1990) and Gromacs† software packages.

A k-medoid clustering (Kaufman & Rousseeuw, 1990) was performed over all structures in the 340 K trajectory projected on the three-dimensional space spanned by the three eigenvectors of the covariance matrix with the largest eigenvalues. For comparison, the same procedure was repeated in the full 60-dimensional space spanned by the 20 backbone atoms in residues 2 to 6. In a k-medoid or k-means clustering, an iterative search is performed for the cluster centers that have a minimal sum of distances to their cluster members. The distribution of transition times between the obtained clusters was compared to a Markov process, which, by definition, is memory-free. Accordingly, deviations between the simulations and the Markov process will point to memory effects. To avoid bias introduced by the unphysically sharp cluster boundaries, the boundary regions were excluded from the analysis and only the "core" parts of each cluster were considered to calculate transition rates. The core of a cluster was defined as that part of each cluster that had a Euclidian distance to the cluster center smaller than the average distance for all cluster members to the cluster center. In addition, to remove short-time oscillations, transitions were taken into account only if the new state was populated for at least 2.5 ps.

To compute the Markov transition rates that best match the transition time distribution obtained from the simulation, a maximum-likelihood approach was employed. Accordingly, the (conditional) probability that the observed frequencies n_{pq} for transitions of state p to q (or that the system remains in or returns to state p if p equals q) are the result of a Markov process with transition rates r_{pq} were estimated from the (conditional) probability to obtain the observed frequencies n_{pq} given the Markov rates r_{pq}, and subsequently maximised. Here we have neglected the influence of the (unknown) *a priori* distribution of Markov processes.

The maximum likelihood is obtained for Markov rates r_{pq} that satisfy:

$$r_{pq} = n_{pq} / \sum_{j=1}^{N} \frac{n_{jj}}{1 - \sum_{i=1, i\neq j}^{N} r_{ij}} \qquad (1)$$

and, therefore, can be computed through iteration. Here N represents the total number of states. Initially, the r_{ij} values were estimated from the distibution of transition frequencies n_{pq}, and subsequently, equation (1) was iterated. The values for r_{pq} were found to converge within a few steps.

Transition frequencies at different time-scales were calculated from the MD trajectory at 340 K and compared to those obtained from a number of simulated true Markov processes constructed from the calculated transition rates.

Memory effects were investigated by monitoring conformational transitions without prior classification of all conformations to clusters. This method has the advantage that it is model-free, in the sense that it is independent of the applied clustering algorithm. The idea is to investigate the path-dependency of conformational transitions. For this purpose, the MD trajectories were scanned for conformational transitions in which the system returns to a conformational state that had been visited immediately before. Those history-biased transition times were compared to unbiased average transition times of the same conformational change, i.e. to those determined without the condition of the system returning to the same state. In accordance to previous studies of the same system (Daura *et al.*, 1999b), a conformational state was defined as all conformations within 1 Å RMSD calculated over the central part of the backbone of the peptide. Conformational transitions were counted if the backbone RMSD between two structures exceeded 2 Å.

Approximate relative configurational Gibbs free energies G_i of phase space regions in the direct vicinity of conformation i were calculated from local phase space densities ρ_i. These densities were estimated from averaged nearest-neighbor phase space distances $\langle d_i \rangle$:

$$G_i = -kT \ln \rho_i; \text{ with } \rho_i \propto \frac{1}{\langle d \rangle_i^m}$$

with the exponent m being the effective dimensionality of the (sub)space in which the densities were determined‡.

Nearest-neighbor phase space distances were determined for all sampled configurations i for each simulation from the average of the distances d_{ij} to the 100 nearest neighbors j (calculated as RMSD values for the 20 backbone atoms in residues 2 to 6 after a least-squares fit). A weighting scheme was applied using a Gaussian function with width d_{i50} chosen as the distance to the 50th nearest neighbor:

$$\langle d \rangle_i = \frac{\sum_{j=1}^{100} d_{ij} e^{-\frac{1}{2}\left(\frac{d_{ij}}{d_{i50}}\right)^2}}{\sum_{j=1}^{100} e^{-\frac{1}{2}\left(\frac{d_{ij}}{d_{i50}}\right)^2}}$$

As a measure for the subspace overlap O_{NM} between N eigenvectors μ_i from one set and M vectors ν_j from another (where the eigenvector sets can have been

† D. Van der Spoel; H. J. C. Berendsen, A. R. van Buuren, E. Apol, P. J. Meulenhoff, A. L. T. M. Sijbers and R. Van Drunen (1995). Gromacs User Manual, available online at http://md.chem.rug.nl/~gmx

‡ Since this effective dimensionality cannot be determined easily, obtained free energies are known only up to a constant factor and thus no units can be given.

obtained from different simulations, or from different simulation fragments, or from selected snapshots or phase space regions) the summed squared inner products (De Groot *et al.*, 1996b) were calculated:

$$O_{NM} = \sum_{i=1}^{N}\sum_{j=1}^{M}(\mu_i \cdot \nu_j)^2$$

yielding zero for two orthogonal spaces and one (for $M \geq N$) for identical sets of vectors. Since two sets of eigenvectors obtained from the same molecular system span the same space, this measure of the overlap always approaches 1 when M approaches the number of degrees of freedom in the system. N is a small number indicating the dimensionality of the subspace that accounts for the majority of all atomic fluctuations, also referred to as the essential subspace (Amadei *et al.*, 1993). A value of 3 was chosen for N for the analysis presented here, since more than 69% of the backbone positional fluctuations are described by the three principal eigenvectors.

HBN energies were evaluated using the hb2net module (Hooft *et al.*, 1996) of the WHAT IF software package (Vriend, 1990). Output "energies" of this knowledge-based method are in units of ideal hydrogen bonds, as derived empirically from a small-molecule structure database.

Acknowledgments

We thank Gert Vriend for assistance in using the WHAT IF program and Daan van Aalten for proofreading the manuscript. B.d.G. was supported by the BIOTECH program of the EU, grant no. BIO4-CT98-0024.

References

Abseher, R. & Nilges, M. (2000). Efficient sampling in collective coordinate space. *Proteins: Struct. Funct. Genet.* **39**, 82-88.

Abseher, R., Horstink, L., Hilbers, C. W. & Nilges, M. (1998). Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. *Proteins: Struct. Funct. Genet.* **31**, 370-382.

Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. (1993). Essential dynamics of proteins. *Proteins: Struct. Funct. Genet.* **17**, 412-425.

Amadei, A., Linssen, A. B. M., De Groot, B. L., Van Aalten, D. M. F. & Berendsen, H. J. C. (1996). An efficient method for sampling the essential subspace of proteins. *J. Biomol. Struct. Dynam.* **13**, 615-626.

Amadei, A., Ceruso, M. A. & Nola, A. D. (1999a). On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Struct. Funct. Genet.* **36**, 419-424.

Amadei, A., de Groot, B. L., Ceruso, M. A., Paci, M., Nola, A. D. & Berendsen, H. J. C. (1999b). A kinetic model for the internal motions of proteins: diffusion between multiple harmonic wells. *Proteins: Struct. Funct. Genet.* **35**, 283-292.

Balsera, M. A., Wriggers, W., Oono, Y. & Schulten, K. (1996). Principal component analysis and long time protein dynamics. *J. Phys. Chem.* **100**, 2567-2572.

Bonvin, A. M. J. J. & van Gunsteren, W. F. (2000). β-Hairpin stability and folding: molecular dynamics studies of the first beta-hairpin of tendamistat. *J. Mol. Biol.* **296**, 255-268.

Chipot, C., Maigret, B. & Pohorille, A. (1999). Early events in the folding of an amphipathic peptide: a multinanosecond molecular dynamics study. *Proteins: Struct. Funct. Genet.* **36**, 383-399.

Clarage, J. B., Romo, T., Andrews, B. K., Pettitt, B. M. & Phillips, G. N., Jr (1995). A sampling problem in molecular dynamics simulations of macromolecules. *Proc. Natl Acad. Sci. USA*, **92**, 3288-3292.

Daggett, V. (2000). Long timescale simulations. *Curr. Opin. Struct. Biol.* **10**, 160-164.

Daura, X., van Gunsteren, W. F., Rigo, D., Jaun, B. & Seebach, D. (1997). Studying the stability of a helical β-heptapeptide by molecular dynamics simulations. *Chem. Eur. J.* **3**, 1410-1417.

Daura, X., Jaun, B., Seebach, D., van Gunsteren, W. F. & Mark, A. E. (1998). Reversible peptide folding in solution by molecular dynamics simulation. *J. Mol. Biol.* **280**, 925-932.

Daura, X., Gademann, K., Jaun, B., Seebach, D., van Gunsteren, W. F. & Mark, A. E. (1999a). Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed.* **38**, 236-240.

Daura, X., van Gunsteren, W. F. & Mark, A. E. (1999b). Folding-unfolding thermodynamics of a β-heptapeptide from equilibrium simulations. *Proteins: Struct. Funct. Genet.* **34**, 269-280.

De Groot, B. L., Amadei, A., Scheek, R. M., Van Nuland, N. A. J. & Berendsen, H. J. C. (1996a). An extended sampling of the configurational space of HPr from *E. coli*. *Proteins: Struct. Funct. Genet.* **26**, 314-322.

De Groot, B. L., Van Aalten, D. M. F., Amadei, A. & Berendsen, H. J. C. (1996b). The consistency of large concerted motions in proteins in molecular dynamics simulations. *Biophys. J.* **71**, 1707-1713.

De Groot, B. L., Hayward, S., Van Aalten, D. M. F., Amadei, A. & Berendsen, H. J. C. (1998). Domain motions in bacteriophage T4 lysozyme; a comparison between molecular dynamics and crystallographic data. *Proteins: Struct. Funct. Genet.* **31**, 116-127.

Duan, Y. & Kollman, P. A. (1998). Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, **282**, 740-744.

Esnouf, R. M. (1997). An extensively modified version of molscript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.* **15**, 132-134, 112-113.

García, A. E. (1992). Large-amplitude nonlinear motions in proteins. *Phys. Rev. Letters*, **68**, 2696-2699.

Grubmüller, H. (1995). Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E*, **52**, 2893-2906.

Hayward, S., Kitao, A. & Gō, N. (1995). Harmonicity and anharmonicity in protein dynamics: a normal modes and principal component analysis. *Proteins: Struct. Funct. Genet.* **23**, 177-186.

Hooft, R. W., Sander, C. & Vriend, G. (1996). Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins: Struct. Funct. Genet.* **26**, 363-376.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, New York.

Kitao, A., Hayward, S. & Gō, N. (1998). Energy landscape of a native protein: jumping-among-minima model. *Proteins: Struct. Funct. Genet.* **33**, 496-517.

Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Merritt, E. A. & Bacon, D. J. (1997). Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505-524.

Nardiu, F., Kemmink, J., Sattler, M. & Wade, R. C. (2000). The *cis*proline($i-1$)-aromatic(i) interaction: folding of the Ala-*cis*Pro-Tyr peptide characterized by NMR and theoretical approaches. *J. Biomol. NMR*, **17**, 63-77.

Pande, V. S. & Rokhsar, D. S. (1999). Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. *Proc. Natl Acad. Sci. USA*, **96**, 9062-9067.

Roccatano, D., Amadei, A., Di Nola, A. & Berendsen, H. J. C. (1999). A molecular dynamics study of the 41-56 beta-hairpin from B1 domain of protein G. *Protein Sci.* **8**, 2130-2143.

Seebach, D., Schreiber, J. V., Abele, S., Daura, X. & van Gunsteren, W. F. (2000). Structure and conformation of β-oligopeptide derivatives with simple proteinogenic side-chains. Circular dichroism and molecular dynamics investigations. *Helv. Chim. Acta*, **83**, 34-57.

Van Aalten, D. M. F., Amadei, A., Vriend, G., Linssen, A. B. M., Venema, G., Berendsen, H. J. C. & Eijsink, V. G. H. (1995a). The essential dynamics of thermolysin: confirmation of hinge-bending motion and comparison of simulations in vacuum and water. *Proteins: Struct. Funct. Genet.* **22**, 45-54.

Van Aalten, D. M. F., Findlay, J. B. C., Amadei, A. & Berendsen, H. J. C. (1995b). Essential dynamics of the cellular retinol binding protein: evidence for ligand induced conformational changes. *Protein Eng.* **8**, 1129-1136.

Van Aalten, D. M. F., Conn, D. A., De Groot, B. L., Findlay, J. B. C., Berendsen, H. J. C. & Amadei, A. (1997). Protein dynamics derived from clusters of crystal structures. *Biophys. J.* **73**, 2891-2896.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52-56.

Edited by R. Huber

(Received 12 December 2000; received in revised form 21 March 2001; accepted 21 March 2001)