Systematic determination of order parameters for chain dynamics using diffusion maps

Andrew L. Ferguson[a], Athanassios Z. Panagiotopoulos[a], Pablo G. Debenedetti[a,1], and Ioannis G. Kevrekidis[a,b]

[a]Department of Chemical and Biological Engineering, and [b]Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544

Edited by Michael L. Klein, Temple University, Philadelphia, PA, and approved June 11, 2010 (received for review March 15, 2010)

We employ the diffusion map approach as a nonlinear dimensionality reduction technique to extract a dynamically relevant, low-dimensional description of *n*-alkane chains in the ideal-gas phase and in aqueous solution. In the case of C_8 we find the dynamics to be governed by torsional motions. For C_{16} and C_{24} we extract three global order parameters with which we characterize the fundamental dynamics, and determine that the low free-energy pathway of globular collapse proceeds by a "kink and slide" mechanism, whereby a bend near the end of the linear chain migrates toward the middle to form a hairpin and, ultimately, a coiled helix. The low-dimensional representation is subtly perturbed in the solvated phase relative to the ideal gas, and its geometric structure is conserved between C_{16} and C_{24}. The methodology is directly extensible to biomolecular self-assembly processes, such as protein folding.

It has long been suspected that cooperative couplings between degrees of freedom render the effective dimensionality of biophysical systems far less than the $3R$-dimensional coordinate space of the R constituent atoms (1–5). This has been framed in the projection operator formalism (6) as a separation of time scales in which the important dynamics reside in a "slow subspace" (7) and is associated with a smooth underlying free energy surface (8). For example, two-dimensional descriptions have been formulated for dialanine (9) and a coarse-grained model of the src homology 3 domain (5).

Calculation of the effective dimensionality of a dynamical system, and identification of order parameters describing the low-dimensional "intrinsic manifold" to which the system dynamics are effectively restrained, is a long-standing problem in as seemingly disparate fields as data visualization (10), speech recognition (11), semisupervised learning (12), and spectral clustering (13). The fraction of native contacts (Q) (8, 14) and the folding probability (P_{fold}) (8, 15) have been used as reaction coordinates for protein folding, but such coarse variables may lump together structurally and kinetically disparate conformations and can prove inadequate for larger proteins with frustrated folding funnels (5, 8). Empirical order parameters also tend to perform poorly on landscapes exhibiting multiple local free-energy (FE) minima or lacking well-defined unfolded and folded basins. Principal components analysis (PCA) is a popular linear dimensionality reduction technique applied extensively to biophysical systems (1–4, 16) which seeks to describe the "essential subspace" (2) of the dynamics by a set of orthogonal vectors oriented along the directions of largest variance in the data. For the highly nonlinear intrinsic manifolds one expects for complex molecular systems (5), the linearity of this technique renders it appropriate in local regions, but results in a poor characterization of the global features (5, 17). This deficiency leads to poor PCA estimates of the effective dimensionality (17) far in excess of the dimensionality of the phase space dynamics determined by Lyapunov analysis (3).

A number of nonlinear dimensionality reduction techniques have emerged in recent years such as Isomap (17), local linear embedding (LLE) (18), and diffusion maps (19–21), which seek to reconstruct the intrinsic manifold by integrating local structural information dictated by the data geometry into a unified global description. Isomap (5, 22) and LLE (23) have been successfully applied to peptide systems, and, although diffusion maps have been used to study phenomena as diverse as chemical reaction networks (24) and defect mobility at an interface (25), they have not been previously applied to systems of biophysical significance.

Path-based techniques such as the finite temperature string (26), nudged elastic band (27), and transition path sampling (28) aim to determine minimum (free) energy routes between metastable states of biophysical systems, from which order parameters and mechanisms may be inferred. Diffusion maps may complement such approaches by furnishing order parameters with which to better characterize metastable basins or by providing a low-dimensional description of the pathway ensemble. Although we were unable to find comparative studies, application of the diffusion map approach to the simulation trajectories in this work was computationally inexpensive, requiring less than 20 h on a single 2.66 GHz processor. Path-based techniques would be more appropriate for systems exhibiting high free-energy barriers, because the diffusion map requires dense sampling of phase space.

N-alkanes are relatively simple molecules whose dynamic and structural behavior nevertheless remains rich and far from well understood, with recent work demonstrating the exotic chain conformations adopted by these molecules (29). Despite the absence of specific interactions or a unique native structure, the behavior of single *n*-alkane chains in water has long been of interest for understanding the role of hydrophobicity in protein folding (30–35). In this work, we have conducted long molecular dynamics simulations of *n*-octane (C_8), *n*-hexadecane (C_{16}), and *n*-tetracosane (C_{24}) in explicit water, and corresponding ideal-gas phase Monte Carlo simulations in which an isolated chain interacts only with itself. By applying diffusion maps to the simulation data, we extracted the intrinsic manifolds which we determined to be approximately three-dimensional and well conserved between the ideal gas and solvated phases. The simulations serve only as a means to sample a canonical distribution of system configurations; the underlying dynamics of the algorithms do not play a role in the diffusion map approach. The physical interpretation of order parameters identified by the diffusion map is unknown a priori, but can be facilitated by correlation with "intermediary" variables, in this case the principal moments of the *n*-alkane gyration tensor. Structural details were resolved by visualizing system configurations at representative data points. We determined three global order parameters for C_{16} and C_{24} describing the degree of collapse, location of the bend, and the handedness of the chain helicity, and determined the low-FE pathway for globular collapse to proceed by a kink and slide mechanism, whereby a bend near the end of the extended chain migrates

Author contributions: A.L.F., A.Z.P., P.G.D., and I.G.K. designed research; A.L.F. performed research; A.L.F., A.Z.P., P.G.D., and I.G.K. analyzed data; and A.L.F., A.Z.P., P.G.D., and I.G.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: pdebene@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1003293107/-/DCSupplemental.

to the center to form a symmetric hairpin, which subsequently collapses into a helical coil.

Diffusion Map

A molecular simulation trajectory recording the coordinates of all R constituent atoms consists of a set of $3R$-dimensional snapshots. Using the diffusion map approach, we seek to arrange the trajectory in a low-dimensional space such that snapshots that are dynamically proximate (i.e., the system may evolve from one to the other on a short time scale) are situated near one another. In the following description of the method, we strive to present a physically motivated summary, reserving a more mathematical treatment for the *SI Text*.

The first step is to establish the dynamic proximity among all snapshot pairs, where, in the absence of an easily computable dynamic measure, we employ a structural similarity metric, with small values implying close dynamic proximity. We choose the rotationally and translationally minimized rmsd between the coordinates of the n-alkane united atom centers (36), denoting the distance between snapshots snap_i and snap_j as rmsd_{ij}. Although this measure ostensibly discards all solvent degrees of freedom, the solvent influences the simulation trajectory and its effect is “encoded” in the n-alkane configurations sampled.

The pairwise distances are now combined with a Gaussian kernel of bandwidth ϵ. For N snapshots, this transformation yields the N-by-N matrix $\mathbf{A}$ with elements

$$A_{ij} = \exp\left(-\frac{(\text{rmsd}_{ij})^2}{2\epsilon}\right) \qquad i,j = 1,\ldots,N. \qquad [1]$$

This step provides a smooth threshold for the pairwise distances, discarding large and retaining small rmsd_{ij} values, where $\sqrt{\epsilon}$ is a characteristic value below which we consider the similarity metric a meaningful measure of dynamic proximity. Following Grassberger and Procaccia’s use of the correlation dimension as a measure of fractal dimensionality (37), Coifman et al. demonstrated that twice the slope of the linear region of a log–log plot of $\sum_{i,j} A_{ij}$ versus ϵ provides an estimate of the effective dimensionality of the system dynamics and delineates the range of suitable ϵ values (38). An example calculation is provided for solvated C_{16} in Fig. S1.

A diagonal matrix $\mathbf{D}$ is constructed from the row sums of $\mathbf{A}$ and used to construct the N-by-N matrix $\mathbf{M}$,

$$D_{ii} = \sum_{j=1}^{N} A_{ij} \qquad i = 1,\ldots,N, \qquad [2]$$

$$\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}. \qquad [3]$$

The eigenvectors of $\mathbf{M}$ arranged in decreasing eigenvalue order, $\{\vec{\phi}_i\}_{i=1}^{N}$, may be efficiently computed as described in *Materials and Methods*, with $\vec{\phi}_1$ the trivial all-ones vector. The k-dimensional diffusion map is the mapping of the ith simulation snapshot into the ith component of each of the top k nontrivial eigenvectors (19, 21, 39)

$$\text{snap}_i \mapsto (\vec{\phi}_2(i),\vec{\phi}_3(i),\ldots,\vec{\phi}_{k+1}(i)). \qquad [4]$$

For brevity, this mapping will henceforth be referred to simply as the “embedding in the top k eigenvectors.” Dimensionality reduction is achieved by mapping the $3R$-dimensional simulation snapshots into a k-dimensional embedding. Determination of an appropriate k is system dependent, and is addressed in *Results and Discussion*.

Free energy surfaces (FES) may be computed over the diffusion map embeddings using the relation $\beta G(\vec{x}) = -\ln \hat{p}(\vec{x}) + \text{const}$, where $\beta = 1/k_B T$, G is the Gibbs free energy, $\hat{p}(\vec{x})$ is a histogram approximation to the density of snapshots at $\vec{x}$, and $\vec{x}$ is a k-dimensional vector of the eigenvector components.

As discussed in more depth in the *SI Text*, if the system dynamics can be well modeled as a diffusion process, and the structural similarity metric is a good descriptor of microscopic diffusive motions, then the diffusion map embedding possesses two important features. (*i*) The Euclidean distance between two snapshots in the diffusion map embedding corresponds to their *diffusion distance*, which may be regarded as the ease with which the system can evolve from one snapshot to the other (19, 39). Snapshots connected by a large number of short pathways have a small diffusion distance and will be embedded close together. (*ii*) The diffusion map embedding captures the slow dynamical motions of the system, capturing the intrinsic or “slow” manifold. Together, these properties make the diffusion map embeddings of the simulation trajectory dynamically meaningful, because paths traced out over this embedding describe the evolution of the system in its fundamental dynamical motions. Although these assumptions are expected to hold for biophysical systems such as this, in the event that they do not, the identified order parameters remain good variables with which to parametrize the evolution of the system from one state to another.

Results and Discussion

***n*-Octane.** The (fractal) effective dimensionalities of the C_8 system in the ideal gas and solvated phases were estimated as 3.0 and 2.7, respectively, suggesting that embeddings may be constructed in the top three eigenvectors. The ordered principal moments of the C_8 gyration tensor (ξ_1,ξ_2,ξ_3) describe the characteristic length of the chain along each of its three principal axes, providing a convenient measure of elongation or globularity (40) and serving as convenient “intermediaries” in the nontrivial task of assigning physical meaning to the order parameters furnished by the diffusion map (Fig. S2 *A* and *B*). They do not, however, map bijectively with these order parameters and so are not in themselves the “right” variables.

Two-dimensional projections of the three-dimensional embedding of the solvated C_8 system into eigenvectors 2, 3, and 4 are presented in Fig. 1. In Fig. 1*A*, the data points are colored according to ξ_1, which correlates well with eigenvector 2 (evec2). For chain molecules such as n-alkanes, ξ_1 is typically much greater than the other principal moments, and is the dominant contribution to the radius of gyration (R_g), which is related to the principal

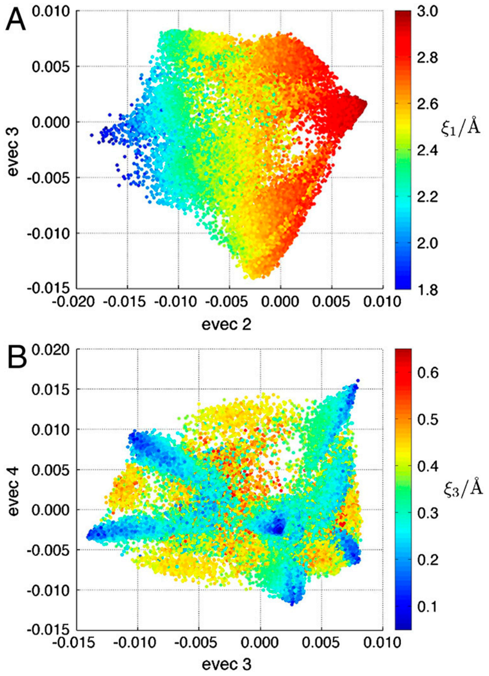

Fig. 1. Two-dimensional elevations of the three-dimensional embedding of the solvated phase C_8 system into evec2, evec3, and evec4. Data points are colored according to the (*A*) first and (*B*) third principal moments of the C_8 gyration tensor.

moments by $R_g^2 = \xi_1^2 + \xi_2^2 + \xi_3^2$. This relationship results in an approximately bijective mapping between ξ_1 and R_g and, accordingly, evec2 also shows good correlation with R_g (Fig. S2 *C* and *D*). In Fig. 1*B*, the points are colored according to ξ_3, which permits the identification of nine distinct low-ξ_3 locales, of which seven are visible as dark blue regions, with the remaining two buried in the point cloud. These regions correspond to local FE minima, all of which are visible as low-lying isosurfaces of the FES in Fig. 2. Visualization of structures (41) residing in each FE well reveals that the basins correspond to various combinations of gauche defects in the chain. The fact that these structures are approximately planar permits their identification with low values of ξ_3.

Figs. S3 and S4 present analogous plots to Figs. 1 and 2 for C_8 in the ideal-gas phase. The fact that the structure of the ideal-gas phase intrinsic manifold and FES are remarkably similar to those in the solvated phase reinforces the assertion that the solvent plays little role in the conformations of short chain *n*-alkanes (34, 42). R_g has previously been suggested by ourselves and others as a good order parameter for *n*-alkane systems (29, 35, 42), and its correlation with the components of the top eigenvector justifies its use as a good one-dimensional descriptor for C_8. However, the diffusion map approach furnishes a more informative three-dimensional embedding in which local minima are separated on the basis of gauche defects, and transitions between them correspond to torsional motions of the chain.

***n*-Hexadecane and *n*-Tetracosane.** The effective dimensionality of ideal gas C_{16} was estimated to lie in the range 3.1–6.2, in agreement with the solvated phase estimate of 3.9–5.6. In the case of C_{24}, the ideal-gas and solvated phase estimates were 3.0–5.4 and 3.2–5.4, respectively. The spread arises from the precise location at which the slope is computed in the $\sum_{i,j} A_{ij}$ vs. ϵ log–log plot (Fig. S1). This ambiguity motivated us to develop an independent measure of the fractal dimensionality of the intrinsic manifold by computing the correlation dimension (37) of the embeddings in successively more eigenvectors, with the value at which this function flattens out, known as the *plateau dimension* (43). Up to 12-dimensional embeddings were constructed to determine a plateau dimension between 2.8–2.9 for the C_{16} systems, and 2.9–3.0 for the C_{24} systems, suggesting that embeddings be constructed in the top three eigenvectors.

The eigenvectors returned by the diffusion map are mutually orthogonal, but embeddings into their components (Eq. **4**) may exhibit functional dependencies. Conceptually, the diffusion map may return multiple order parameters characterizing the same dynamic pathway. To draw an analogy with multivariate Fourier series, $\sin(x)$ and $\sin(2x)$ are independent Fourier components both oriented in the same spatial direction. In the case of solvated C_{16}, such a dependency collapses an embedding into the components of evec2 and evec3 onto an effectively one-dimensional curve (Fig. S5*A*). By fitting two piecewise continuous quartic functions, this curve was parametrized by its scalar valued arclength (Fig. S5*B*), permitting an embedding in [evec2/3 arclength, evec4, evec6], where evec5 exhibited a functional dependency on arclength, and was omitted in favor of evec6. Similarly, ideal gas C_{16} was embedded in [evec2/3 arclength, evec5, evec6], and C_{24} in [evec2/4 arclength, evec3, evec5] in both the ideal-gas and solvated phase.

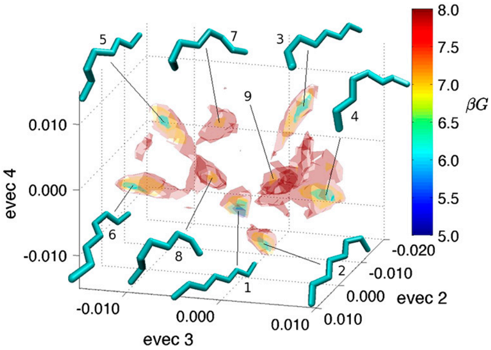

Fig. 2. FES of the solvated phase C_8 system embedded in evec2, evec3, and evec4 with representative chain conformations. As in Figs. 3, 4, and 5*A*, molecules are oriented such that the head is farther from the reader, and solvent has been removed for clarity. The range of βG (where G is the Gibbs free energy, and $\beta = 1/k_B T$) is 3.6–10.3, with isosurfaces plotted at $\beta G = 5$, 6, 7, and 8. The ninth low-ξ_3 region midway between structures 2 and 3 has not been associated with a distinct structure, because it describes transitory conformations between 2 and 3 containing gauche defects in both the head and tail.

Fig. 3 presents elevations of the three-dimensional embedding of the solvated C_{24} system, with data points colored according to the principal moments of the C_{24} gyration tensor to assist in the physical interpretation of the eigenvectors and arclength. Fig. 3*A* demonstrates that ξ_1, which describes the extent of the molecule along its longest axis, is anticorrelated with arclength, permitting arclength to be interpreted as the degree of molecular collapse. Structural details are resolved by visualizing representative chain conformations at increasing values of arclength while holding evec3 = evec5 = 0. The progression of structures 1 → 2 → 3 → 6 in Fig. 3 tracks the collapse from an extended conformation via a bend in the middle of the chain to a tight, symmetric hairpin.

Fig. 3*B* presents the same projection of the manifold as Fig. 3*A*, but with the data points colored according to ξ_2, describing the extent of the chain along its second longest axis. For values of arclength between 0.04 and 0.07, evec5 values around zero correspond to high values of ξ_2, whereas values of evec5 away from

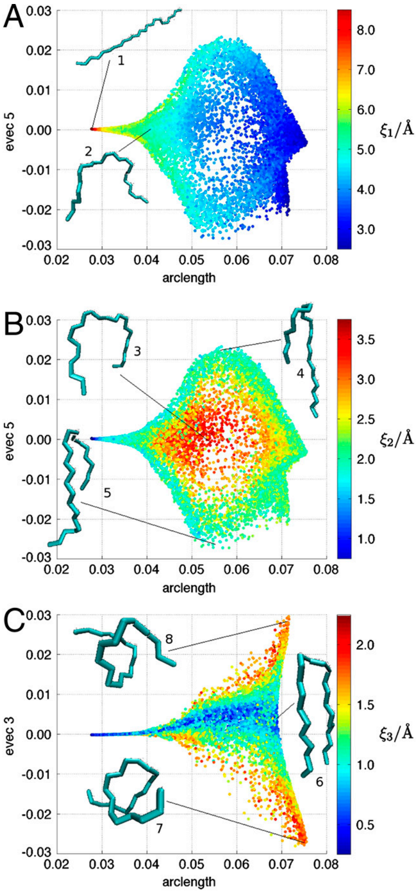

Fig. 3. Two-dimensional elevations of the three-dimensional embedding of the solvated phase C_{24} system in evec2/4 arclength, evec3, and evec5. Data points are colored according to the (*A*) first, (*B*) second, and (*C*) third principal moments of the C_{24} gyration tensor.

zero are associated with lower values of ξ_2. Visualization of representative chain conformations reveals the details of transitions described by evec5. Structure 3 has an evec5 value of around zero, and is a loose hairpin in which the bend is located approximately at the center of the chain, with the looseness reflected in a large value of ξ_2. As evec5 is increased to approach structure 4, the kink migrates toward the tail and is accompanied by a tightening of the hairpin and a reduction in ξ_2. Similarly, as evec5 is reduced from zero to approach structure 5, the hairpin tightens, but the kink now migrates toward the head of the chain. No large ξ_2 values are observed at values of arclength below 0.04 and above 0.07, because structures in the first case correspond to linearly extended conformations (structure 1), and in the second to tight hairpins (structure 6) and coils (structures 7 and 8).

The data points in the elevation of the manifold presented in Fig. 3*C* are colored according to ξ_3, describing the extent of the *n*-alkane chain along its shortest principal axis. Structure 6 has a value of evec3 near zero, and a small value of ξ_3 due to the planarity of this tight, symmetric hairpin. Moving toward positive (negative) values of evec3 corresponds to a transition to a left-handed (right-handed) helical coil depicted by structure 8 (structure 7), which is accompanied by an increase in ξ_3 due to a transition from a planar to a more globular form. Therefore, evec3 may be interpreted as describing the deviations from planarity of the molecule, distinguishing the handedness of such deviations toward right- or left-handed helical coiled structures akin to those observed by Chakrabarty and Bagchi (29).

Pairwise rmsd distances were computed using a consistent definition of the head-tail directionality of the *n*-alkane chains in order to yield a dynamically meaningful similarity metric. Because the chemical structure of *n*-alkanes is, however, identical irrespective of which end is defined as the head, the head-tail symmetry emerged naturally in order parameters extracted by the diffusion map, and is apparent in the approximate planes of symmetry in the structure and coloration of the manifold elevations in Fig. 3 *B* and *C*. For instance, structures containing a kink near the head (structure 5) are related by a head-tail inversion to those with a kink near the tail (structure 4).

The intrinsic manifold of solvated C_{16} (Fig. S6) is remarkably similar to that of solvated C_{24} (Fig. 3), and the intrinsic manifolds of ideal gas C_{16} (Fig. S7) and C_{24} (Fig. S8) exhibit striking similarity to those of the corresponding solvated systems. The similarity of the fundamental structure of the intrinsic manifold in all four systems suggests that the chain conformations explored by C_{16} and C_{24} in the ideal-gas and solvated phases are largely the same (34, 42), with the slow dynamics restrained to similar low-dimensional attractors.

We now turn from a consideration of the structure to an analysis of the conformational population distribution by constructing FES over the manifolds. Fig. 4 *A* and *B* present FES for C_{24} in the ideal-gas and solvated phases, respectively. Fig. 4*A* illustrates the presence of a low-FE "doughnut" encircling a higher FE region, and linking extended and collapsed conformations by two distinct routes. The pathway indicated by the upper arrow illustrates the progression from low to high arclength via positive values of evec5, whereas the lower arrow indicates a route via negative evec5 values. The representative structures projected onto the manifold demonstrate that the transitions proceed by a kink and slide mechanism, where a kink developing near the head (evec5 < 0) or tail (evec5 > 0) of an extended chain migrates toward the middle to form a tight, symmetric hairpin. These pathways follow an FE isosurface and are therefore essentially barrierless. The direct route from low to high arclength with evec5 $= 0$ traverses the "doughnut hole," as indicated by the dashed arrow, and corresponds to a symmetric collapse pathway whereby a loose hairpin closes symmetrically into a tight hairpin. This route is less favored than the asymmetric collapse pathways, containing an extended 1-2*kT* FE barrier. Populations of helical coils at large positive and negative values of evec3 are too low to be resolved on the FES.

Fig. 4*B* demonstrates that the kink and slide mechanism remains the low-FE pathway for chain collapse in the solvated phase, as indicated by the arrows. The symmetric collapse pathway was so rarely sampled throughout the 30 ns simulation that we compute an infinitely high barrier at the resolution of our FES. Accordingly, the low-FE incursion on the low-arclength side of the doughnut hole (indicated by *) corresponds to loose, symmetric hairpins and represents a dead end to chain collapse. The helical coils are significantly more stable in the solvated phase relative to the ideal gas, with similar free energies as the tight, symmetric hairpins. That such structures exist in both the ideal-gas and solvated phases suggest that these morphologies are inherent to the *n*-alkane chain rather than a product of the aqueous environment (29), but are significantly stabilized by the solvent interaction, presumably by the hydrophobic effect (44, 45).

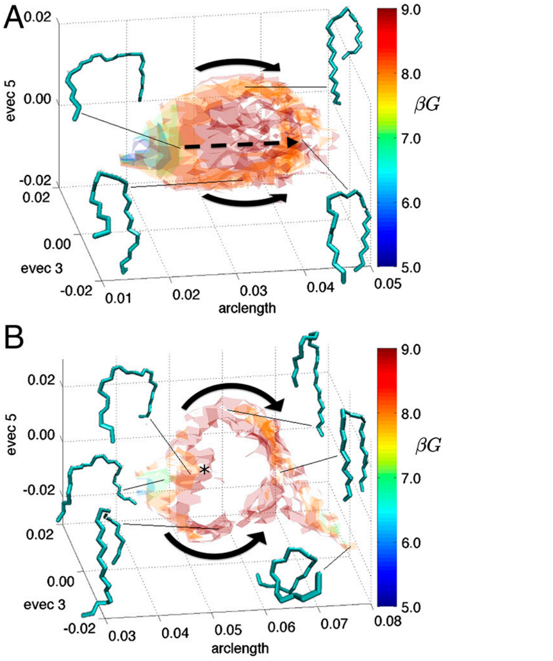

Fig. 4. FES for the C_{24} chain in the (*A*) ideal-gas and (*B*) solvated phase. In both cases the embedding is constructed in evec2/4 arclength, evec3, and evec5. The range of βG is 2.4–10.3 in the ideal-gas phase and 1.6–10.3 in the solvated phase, with isosurfaces plotted at $\beta G = 5$, 6, 7, 8, and 9 in each case. The essentially one-dimensional low-arclength tip apparent in Fig. 3*C* and Fig. S8*C* constitutes the global FE minimum, but is not resolved in the three-dimensional plot. Solid arrows indicate collapse pathways by the kink and slide mechanism, and the dashed arrow in *A* indicates the symmetric hairpin collapse route. The * in *B* marks the low-FE incursion mentioned in the text, and the positive evec3 wing is not visible due to the perspective of the plot.

The major features of the C_{24} ideal-gas and solvated FES are conserved in the case of C_{16} and are presented in Fig. S9. The low-FE pathway to collapse for solvated C_{16} also proceeds by a kink and slide mechanism, followed by further collapse into a helical coil. The symmetric collapse route is as favorable as the asymmetric pathways in the ideal-gas phase, but whereas the asymmetric routes remain barrierless in the solvated phase, the symmetric pathway contains a $\gg kT$ barrier, with precise determination of the height frustrated by inadequate sampling of this region. Depopulation of the symmetric collapse pathway in the solvated phase relative to the ideal gas appears to be the root of our previously reported solvent-induced FE barrier between extended and collapsed *n*-alkane conformations (42). As for C_8, the diffusion map approach has identified ξ_1, or equivalently R_g, as a good one-dimensional order parameter, although this representation is inferior to a three-dimensional description in terms of the degree or chain collapse, position of the bend in the chain, and handedness of the helical coil.

Although an objective validation of the dynamical relevance of the low-dimensional description would require evaluation of the committor probabilities along the collapse pathway (46), the identified variables are preserved when considering partial simulation trajectories (compare Fig. 5*A* and Fig. S6*A*), and appropriately characterize collapse events observed in the trajectories (Movie S1).

Solvent Analysis. The role of the solvent was probed by calculating the solvent-excluded cavity volume occupied by the C_{16} chain using a test probe insertion procedure detailed in *Materials and Methods*. Due to the computational expense of the procedure, a contiguous 3 ns portion of the full 30 ns solvated trajectory was considered. Application of the diffusion map to the partial trajectory proved robust, resulting in a three-dimensional embedding of the intrinsic manifold (Fig. 5*A*) very similar in structure to those for the complete C_{16} and C_{24} trajectories (Fig. S6*A* and Fig. 3). The data points in Fig. 5*A* are colored according to the cavity volume, illustrating that chain collapse from low-arclength, extended conformations to high-arclength, tight, symmetric hairpins and helical coils, is accompanied by an increase in the solvent-excluded cavity volume. Considering arclength $= 0.12$ as a cutoff, the mean cavity volume of the extended conformations is 16 ± 11 Å^3, compared to 39 ± 20 Å^3 for the collapsed structures. Large variances are expected in the measurement of a dynamic void volume depending on the collective motion of many solvent molecules.

Fig. 5*B* provides a three-dimensional view of the solvated C_{24} intrinsic manifold presented in Fig. 3, together with representative snapshots of the *n*-alkane chain and surrounding water molecules, to illustrate the details of the solvent as the chain collapses via the kink and slide mechanism. The interior of the low-arclength, loose hairpin is hydrated (snapshot A in Fig. 5*B*), as is apparent by the presence of water molecules between the arms of the chain. Solvent is excluded from within a bend in the tail of the chain (B), which subsequently slides down to form a tight, symmetric hairpin with a dry interior (C), and collapses further into a helical coil with a dry core (D) (Movie S1). The expulsion of solvent from the chain interior is captured by the increase in the solvent-excluded cavity with increasing arclength and, in this sense, solvent effects are contained within the extracted order parameters without explicit consideration of solvent degrees of freedom.

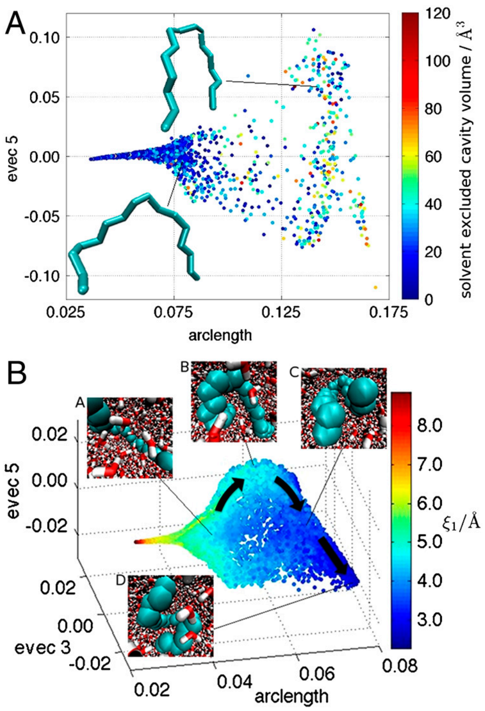

Fig. 5. Solvent analysis. (*A*) Two-dimensional elevation of the three-dimensional embedding of the solvated phase C_{16} system constructed from a 3 ns portion of the full 30 ns trajectory. The embedding is constructed in evec2/3 arclength, evec4, and evec5, after a small rotation of the manifold to present a view consistent with that of Fig. 3. Data points are colored according to the solvent-excluded cavity volume occupied by the C_{16} chain. (*B*) Three-dimensional view of the solvated phase C_{24} system embedded in evec2/4 arclength, evec3, and evec5. Data points are colored according to the first principal moment of the C_{24} gyration tensor.

The depopulation of the symmetric hairpin collapse pathway in the solvated phase relative to the ideal gas (Fig. 4) is apparently due to a wetting/dewetting FE barrier similar to that observed by Lum et al. for bundles of hydrophobic cylinders (47), with the kink and slide mechanism providing a less expensive route to collapse, avoiding the collective expulsion of interior solvent molecules. A study of a hydrophobic 12-mer by Miller et al. (46) determined the low-FE collapse pathway to proceed via a drying transition at a bend in the middle of the chain. Athough the sensitivity of hydrophobic hydration mechanics to subtle differences in the force field (29) may underlie these differences, the large monomers of Miller et al. may permit this model to be interpreted as a coarse-grained representation of ~C_{60} (35), suggesting an alternative collapse mechanism for long chains.

Conclusions

We have demonstrated an application of diffusion maps to systematically recover a small number of "good" order parameters from simulations of C_8, C_{16}, and C_{24} *n*-alkane chains in the ideal-gas and solvated phases. The intrinsic manifolds upon which the dynamics of each system effectively lie were reconstructed by embedding the simulation trajectories into these order parameters, and a physical interpretation of the parameters was facilitated by correlating them with the principal moments of the *n*-alkane gyration tensors. FES constructed on the manifold are dynamically meaningful, with the low-FE pathways providing mechanistic insight. In the case of C_8, the local FE minima were separated on the basis of the dihedral angles of the chain, with transitions between minima corresponding to torsional chain dynamics. The ideal-gas and solvated phase FES were strikingly similar in both structure and depth of the local minima, indicating relatively little effect of the solvent interaction on the conformations of the chain.

For the C_{16} and C_{24} systems, the diffusion map approach identified three global order parameters describing the degree of collapse, location of the bend in the chain, and the handedness of the chain helicity. Although the overall structure of the FES was conserved between the ideal-gas and solvated phases, helical coil conformations were stabilized in the solvated phase relative to the ideal gas, whereas the collapse pathway corresponding to the tightening of a loose symmetric hairpin was destabilized. The low-FE pathway for the collapse of both chains in solvent was observed to proceed first by a kink and slide mechanism, whereby a kink near the end of the chain migrates to the middle to form a symmetric hairpin with a dry interior, followed by further collapse into a helical coil. These results suggest that the FES and underlying dynamical motions of *n*-alkanes of lengths between C_{16} and C_{24} are well conserved, although the extent of this range and the manner in which the short chain behavior merges into this regime remains to be determined.

Materials and Methods

Molecular Simulations. Solvated phase molecular dynamics simulations were conducted with the GROMACS 4.0.2 simulation suite (48) employing the Transferable Potentials for Phase Equilibria potential for the *n*-alkane chains (49) and the simple point charge model of water (50). PRODRG2 (51) assisted in the building of *n*-alkane topologies. Lennard–Jones interactions were smoothly switched to zero at 14 Å, whereas real-space electrostatic interactions were truncated at 15 Å and the reciprocal space treated with Particle Mesh Ewald (52). Systems were subjected to energy minimization, 5 ps of position restrained dynamics, and 1 ns of equilibration, before conducting 30 ns production runs at 298 K and 1 bar maintained by a Nosé–Hoover thermostat (53, 54) and Parrinello–Rahman (55) barostat. Snapshots were saved every 1 ps. Conformationally biased (56), ideal-gas Monte Carlo simulations were conducted for 150,000 steps, saving snapshots every fifth step.

Solvent-Excluded Cavity Volumes. A 0.2-Å cubic mesh was placed over the simulation box and the solvation cavity defined as those cells for which the insertion of a 3.75-Å spherical probe into the center of the cell did not result in overlap with any water O atom centers (42). The probe radius was selected so as to result in a probability of zero overlap insertions in bulk water of less than 10^{-7} (57).

Eigenvector Computation. The top 20 eigenvectors and eigenvalues of $30{,}001 \times 30{,}001$ matrices were computed by the Implicitly Restarted Arnoldi Method implemented in the Parallel ARPACK libraries (58). Matrix storage scales as the square of the number of snapshots, but is independent of ostensible system dimensionality.

ACKNOWLEDGMENTS. Computations were performed at the Terascale Infrastructure for Groundbreaking Research in Engineering and Science facility at Princeton University. P.G.D. acknowledges the National Science Foundation (Collaborative Research in Chemistry Grants CHE-0404699 and CHE-0908265). A.Z.P. acknowledges the Department of Energy (Grant DE-SC-0002128) and the Princeton Center for Complex Materials (Materials Research Science and Engineering Center Grant DMR-0819860). I.G.K. acknowledges the US Department of Energy (Grant DE-PS02-08ER08-13).

1. García AE (1992) Large-amplitude nonlinear motions in proteins. *Phys Rev Lett* 68:2696–2699.
2. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. *Proteins* 17:412–425.
3. Hegger R, Altis A, Nguyen PH, Stock G (2007) How complex is the dynamics of peptide folding? *Phys Rev Lett* 98:028102–4.
4. Zhuravlev PI, Materese CK, Papoian GA (2009) Deconstructing the native state: Energy landscapes, function, and dynamics of globular proteins. *J Phys Chem B* 113:8800–8812.
5. Das P, Moll M, Stamati H, Kavraki LE, Clementi C (2006) Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc Natl Acad Sci USA* 103:9885–9890.
6. Zwanzig R (2001) *Nonequilibrium Statistical Mechanics* (Oxford Univ Press, New York), pp 143–168.
7. Hummer G, Kevrekidis IG (2003) Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J Chem Phys* 118:10762–10773.
8. Cho SS, Levy Y, Wolynes PG (2006) P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc Natl Acad Sci USA* 103:586–591.
9. Bolhuis PG, Dellago C, Chandler D (2000) Reaction coordinates of biomolecular isomerization. *Proc Natl Acad Sci USA* 97:5877–5882.
10. Buja A, et al. (2008) Data visualization with multidimensional scaling. *J Comput Graph Stat* 17:444–472.
11. Elman JL, Zipser D (1988) Learning the hidden structure of speech. *J Acoust Soc Am* 83:1615–1626.
12. Zhu X, Ghahramani Z, Lafferty J (2003) *Proceedings of the 20th International Conference on Machine Learning*, eds T Fawcett and N Mishra (AAAI Press, Washington),–912–919.
13. Zelnik-Manor L, Perona P (2004) *Advances in Neural Information Processing Systems 17*, eds LK Saul, Y Weiss, and L Bottou (MIT Press, Cambridge, MA), pp 1601–1608.
14. Best RB, Hummer G (2005) Reaction coordinates and rates from transition paths. *Proc Natl Acad Sci USA* 102:6732–6737.
15. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES (1998) On the transition coordinate for protein folding. *J Chem Phys* 108:334–350.
16. Sanbonmatsu KY, García AE (2002) Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins* 46:225–234.
17. Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290:2319–2323.
18. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290:2323–2326.
19. Coifman RR, et al. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc Natl Acad Sci USA* 102:7426–7431.
20. Coifman RR, Lafon S (2006) Diffusion maps. *Appl Comput Harmon Anal* 21:5–30.
21. Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput* 15:1373–1396.
22. Plaku E, Stamati H, Clementi C, Kavraki LE (2007) Fast and reliable analysis of molecular motion using proximity relations and dimensionality reduction. *Proteins* 67:897–907.
23. Kentsis A, Gindin T, Mezei M, Osman R (2007) Calculation of the free energy and cooperativity of protein folding. *PLoS One* 2:e446.
24. Singer A, Erban R, Kevrekidis IG, Coifman RR (2009) Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proc Natl Acad Sci USA* 106:16090–16095.
25. Sonday BE, Haataja M, Kevrekidis IG (2009) Coarse-graining the dynamics of a driven interface in the presence of mobile impurities: Effective description via diffusion maps. *Phys Rev E* 80:031102–031111.
26. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G (2006) String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J Chem Phys* 125:024106–024115.
27. Jónsson H, Mills G, Jacobsen KW (1998) *Classical and Quantum Dynamics in Condensed Phase Simulations*, eds BJ Berne, G Ciccoti, and DF Coker (World Scientific, Singapore), pp 385–404.
28. Bolhuis PG, Chandler D, Dellago C, Geissler PL (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu Rev Phys Chem* 53:291–318.
29. Chakrabarty S, Bagchi B (2009) Self-organization of n-alkane chains in water: Length dependent crossover from helix and toroid to molten globule. *J Phys Chem B* 113:8446–8448.
30. Tanford C (1972) Hydrophobic free energy, micelle formation and the association of proteins with amphiphiles. *J Mol Biol* 67:59–74.
31. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. *Adv Protein Chem* 14:1–63.
32. Mountain RD, Thirumalai D (2003) Molecular dynamics simulations of end-to-end contact formation in hydrocarbon chains in water and aqueous urea solution. *J Am Chem Soc* 125:1950–1957.
33. Huang DM, Chandler D (2002) The hydrophobic effect and the influence of solute-solvent attractions. *J Phys Chem B* 106:2047–2053.
34. Sun L, Siepmann JI, Schure MR (2006) Conformation and solvation structure for an isolated n-octadecane chain in water, methanol, and their mixtures. *J Phys Chem B* 110:10519–10525.
35. Athawale MV, Goel G, Ghosh T, Truskett TM, Garde S (2007) Effects of lengthscales and attractions on the collapse of hydrophobic polymers in water. *Proc Natl Acad Sci USA* 104:733–738.
36. Maiorov VN, Crippen GM (1995) Size-independent comparison of protein three-dimensional structures. *Proteins* 22:273–283.
37. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. *Physica D* 9:189–208.
38. Coifman RR, Shkolnisky Y, Sigworth FJ, Singer A (2008) Graph laplacian tomography from unknown random projections. *IEEE T Image Process* 17:1891–1899.
39. Nadler B, Lafon S, Coifman RR, Kevrekidis I (2006) *Advances in Neural Information Processing Systems 18*, eds Y Weiss, B Schölkopf, and J Platt (MIT Press, Cambridge, MA), pp 955–962.
40. Theodorou DN, Suter UW (1985) Shape of unperturbed linear polymers: Polypropylene. *Macromolecules* 18:1206–1214.
41. Humphrey W, Dalke A, Schulten K (1996) VMD—visual molecular dynamics. *J Mol Graphics* 14:33–38.
42. Ferguson AL, Debenedetti PG, Panagiotopoulos AZ (2009) Solubility and molecular conformations of n-alkane chains in water. *J Phys Chem B* 113:6405–6414.
43. Sauer T, Yorke JA, Casdagli M (1991) Embedology. *J Stat Phys* 65:579–616.
44. Widom B, Bhimalapuram P, Koga K (2003) The hydrophobic effect. *Phys Chem Chem Phys* 5:3085–3093.
45. Chandler D (2005) Interfaces and the driving force of hydrophobic assembly. *Nature* 437:640–647.
46. Miller TF, Vanden-Eijnden E, Chandler D (2007) Solvent coarse-graining and the string method applied to the hydrophobic collapse of a hydrated chain. *Proc Natl Acad Sci USA* 104:14559–14564.
47. Lum K, Chandler D, Weeks JD (1999) Hydrophobicity at small and large length scales. *J Phys Chem B* 103:4570–4577.
48. van der Spoel D, et al. (2005) Gromacs: Fast, flexible, and free. *J Comput Chem* 26:1701–1718.
49. Martin MG, Siepmann JI (1998) Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. *J Phys Chem B* 102:2569–2577.
50. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans JPullman B (1981) *Intermolecular Forces* (Reidel, Dordrecht, The Netherlands), pp 331–342.
51. Schüttelkopf A, van Aalten D (2004) PRODRG: A tool for high throughput crystallography of protein-ligand complexes. *Acta Crystallogr D* 60:1355–1363.
52. Essmann U, et al. (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–8593.
53. Nosé S (1984) A unified formulation of the constant temperature molecular dynamics methods. *J Chem Phys* 81:511–519.
54. Hoover WG (1985) Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A* 31:1695–1697.
55. Parinello M, Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* 52:7182–7190.
56. Frenkel D, Smit B (2002) *Understanding Molecular Simulation: From Algorithms to Applications* (Academic, San Diego), 2nd Ed, pp 331–353.
57. Hummer G, Garde S, Paulitis M, Pratt L (1998) Hydrophobic effects on a molecular scale. *J Phys Chem B* 102:10469–10482.
58. Maschhoff KJ, Sorensen DC (1996) *Proceedings of the Third International Workshop on Applied Parallel Computing, Industrial Computation and Optimization*, eds J Wasniewski, J Dongarra, K Madsen, and D Olesen (Springer, London), pp 478–486.