How Complex Is the Dynamics of Peptide Folding?

Rainer Hegger,* Alexandros Altis, Phuong H. Nguyen, and Gerhard Stock

Institute of Physical and Theoretical Chemistry, J. W. Goethe University, Max-von-Laue-Strasse 7, 60438 Frankfurt, Germany

(Received 13 September 2006; published 12 January 2007)

Classical molecular dynamics simulations of the folding of alanine peptides in aqueous solution are analyzed by constructing a deterministic model of the dynamics, using methods from nonlinear time series analysis. While the dimension of the free energy landscape increases with system size, a Lyapunov analysis shows that the effective dimension of the dynamic system is rather small and even decreases with chain length. The observed reduction of phase space is a nonlinear cooperative effect that is caused by intramolecular hydrogen bonds that stabilize the secondary structure of the peptides.

DOI: 10.1103/PhysRevLett.98.028102 PACS numbers: 87.15.He, 05.45.Tp

The complexity of a dynamic system represents a well-established concept in the theory of nonlinear dynamics. It is often associated with the fact that the "effective dimension" of the system [1], that is, the dimension of the subspace a trajectory $\vec{x}(t) \in \mathbb{R}^n$ will occupy in the course of its time evolution $\dot{\vec{x}}(t) = \vec{f}(\vec{x}(t))$, can be much smaller than n, the dimension the problem is formulated in. This dimensionality reduction is caused by nonlinear couplings which give rise to cooperative or synchronization effects and consequently reduce the effective number of degrees of freedom. The significance of the concept becomes apparent in the case of a dissipative chaotic system, whose effective dimension typically is a noninteger number. Apart from its conceptional value, the effective dimension of a dynamic system is of practical interest since it may be calculated from measured or simulated data, e.g., by estimating the correlation dimension [2] or the Lyapunov exponents from which the Kaplan-Yorke dimension [3] may be obtained.

In this Letter, we wish to apply the concept of dimensionality to the interpretation of classical molecular dynamics (MD) simulations [4]. While MD simulations describe biomolecular processes such as folding and molecular recognition in atomistic detail (i.e., $3N$-6 coordinates for an N-atomic system), it is clear that the many geometrical constraints of the molecule (e.g., covalent and hydrogen bonds) result in a considerable reduction of the effective number of degrees of freedom. In practice, the structural dynamics of biomolecules is often described in terms of the molecule's free energy surface (the "energy landscape" [5,6]), which is represented as a function of empirically introduced "reaction coordinates" (e.g., the fraction of native contacts formed or the root mean squared distance to the native structure). Alternatively, one may employ a principal component (PC) analysis of the trajectory [7,8], that is, a linear transformation of the coordinate system such that the instantaneous linear correlations between the variables are removed. Ordering the eigenvalues of the transformation decreasingly, it has been shown that a large part of the system's fluctuations can be described in terms of only the first few PCs, which may serve as reaction coordinates. While these coordinates in some sense represent the essential dynamics of the system [9], in general it is not clear how to determine the effective dimension of a biomolecular MD simulation, since there always is some ambiguity in the choice of the reaction coordinates. As a first attempt to assess the complexity of a biomolecular system, in this work we (i) perform MD simulations of various peptide systems and extract time series that account for their structural dynamics, (ii) construct a deterministic model of the dynamics using methods from nonlinear time series analysis, and (iii) perform a Lyapunov analysis to calculate their effective dimension.

As molecular systems we have chosen the alanine peptides Ala$_n$ with n = 3, 5, 7, and 10 in aqueous solution (see Fig. 1), for which 100 ns MD simulations at 300 K were

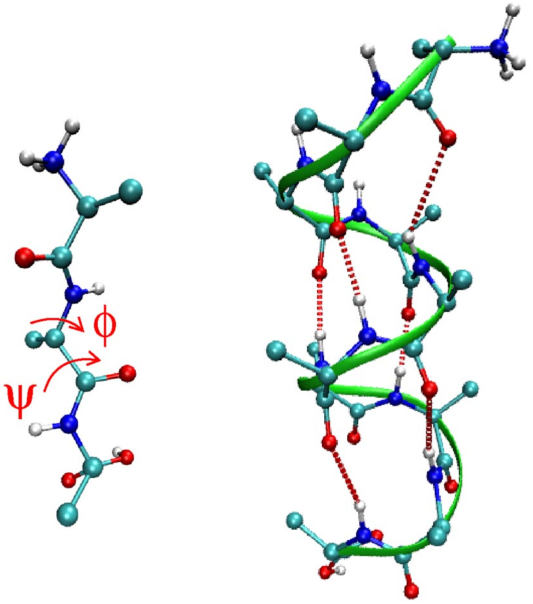

FIG. 1 (color online). MD snapshots of (left) an extended conformation of Ala$_3$ showing the central backbone dihedral angles ϕ and ψ, and (right) the α_R helix conformation of Ala$_{10}$ indicating the stabilizing $n-(n+4)$ hydrogen bonds.

performed using the GROMACS program suite [10], the GROMOS96 force field 43a1 [11], and the simple point charge water model [12] (for details see Ref. [13]). Unlike to proteins, these systems are too small to adopt a stable native structure, but exhibit reversible folding and unfolding of their secondary structure. Since this large amplitude motion results in a strong mixing of internal and global motion (while only the internal motion is of interest), we choose internal coordinates to describe the peptide structure, i.e., their (ϕ_k, ψ_k) backbone dihedral angles ($k = 1, \ldots, n-2$), see Fig. 1. To circumvent problems associated with the fact that angles are circular variables, the angles are mapped onto a Cartesian-like space via $x_{4k} = \cos\phi_k$, $x_{4k-1} = \sin\phi_k$, $x_{4k-2} = \cos\psi_k$, and $x_{4k-3} = \sin\psi_k$, resulting in $4(n-2)$ variables [13]. To remove linear correlations, a PC analysis of the MD trajectory $\vec{x}(t)$ is performed, yielding the PC analysis eigenvectors $\vec{u}_i$ and the corresponding PCs $v_i(t) = \vec{x}(t) \cdot \vec{u}_i$, which serve as a time series for the subsequent analysis.

As a first example, Fig. 2 shows the time series $v_i(t)$, the distributions $P(v_i)$, and the autocorrelation functions $C_i(t)$ obtained for the first two PCs of the Ala$_3$ system. Both distributions exhibit multiple peaks which correspond to different conformational states of the peptide. For the first component, the peak at $v_1 \approx -1.7$ reflects the right-handed helix conformation α_R, while the peak at $v_1 \approx 0.2$ reflects extended conformations of the peptide (see Fig. 1). Invoking the second PC, the latter can be decomposed in the poly-L-proline II (P_{II}) conformation and the fully extended (β) conformation [14]. The transitions between these states occur on a 200 ps ($\alpha_R \leftrightarrow \beta$) and 20 ps ($P_{\text{II}} \leftrightarrow \beta$) time scale, respectively. While the three conformational states of Ala$_3$ can be described using only two PCs, the situation is more involved for the longer peptides. For the Ala$_{10}$ system, for example, Fig. 3 shows that the distributions of the first two PCs are characterized by a prominent double peak corresponding to an α_R-type folded state (see Fig. 1), and a large range of extended and intermediate states corresponding to unfolded structures of the peptide. To discriminate these states, in total eight PCs are required. An analysis of the time evolution of the first PC reveals collective conformational transitions, accounting for the reversible folding and unfolding of the secondary structure of the peptide.

Performing a PC analysis of a MD trajectory, only the distribution of the first, say d_{EL}, PCs exhibit multiple peaks, while the remaining distributions $P(v_i)$ with $i > d_{\text{EL}}$ are single peaked and approach a Gaussian shape with increasing i [9]. That is, the distributions $P(v_i)$ with $i > d_{\text{EL}}$ describe the fluctuations of the peptide within a specific conformational state, while the distributions with $i \leq d_{\text{EL}}$ define these conformation states. Hence d_{EL} can be considered as the dimension of the free energy landscape $\Delta G(\{v_i\}) \propto -\ln P(\{v_i\})$, because ΔG shows nontrivial structure only along the first d_{EL} PCs. As listed in Table I, d_{EL} increases with system size, i.e., from 2 for Ala$_3$ to 8 for Ala$_{10}$.

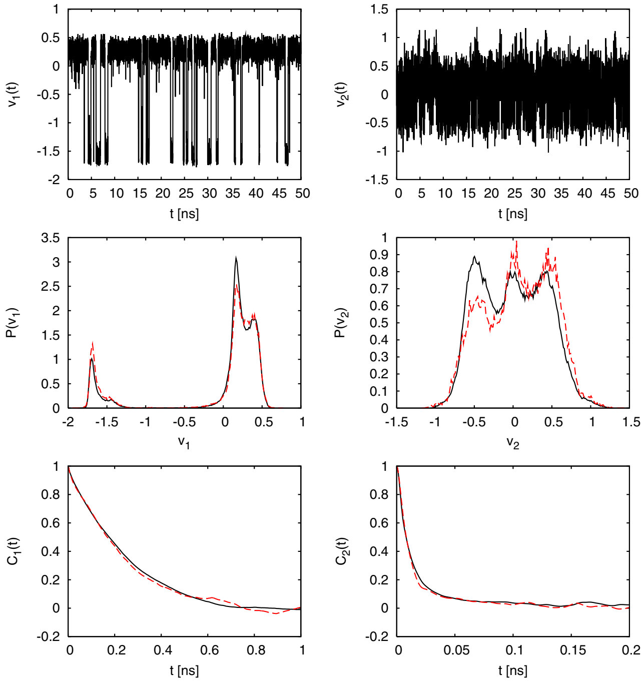

FIG. 2 (color online). Time series $v_i(t)$, distributions $P(v_i)$, and autocorrelation functions $C_i(t)$ obtained for the first two principal components of the Ala$_3$ system. The solid black lines represent the results of the MD simulation, the dashed red (or gray) lines correspond to results from the nonlinear model of the dynamics.

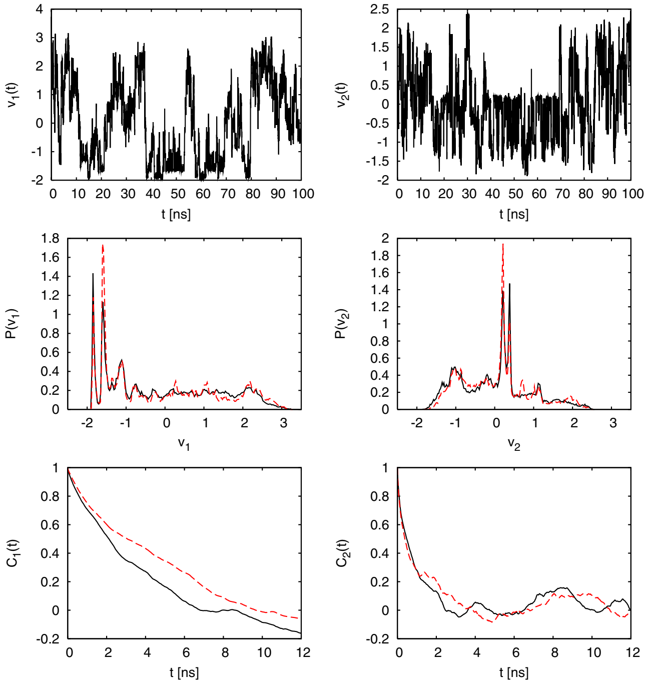

FIG. 3 (color online). Same as in Fig. 2, but for the Ala$_{10}$ system.

It should be emphasized, however, that the energy landscape dimension d_{EL} is conceptually different from the effective dimension of the dynamics in phase space. In principle, the latter can be obtained directly from a Lyapunov analysis of the MD trajectory [15]. In practice, though, the ubiquitous noise on the data prevents an accurate calculation of the Lyapunov exponents, which account for the sensitivity of the trajectory with respect to infinitesimal small deviations of its initial conditions. To overcome this problem, we employ the methods of nonlinear time series analysis [16] and construct a deterministic model of the dynamics which reproduces the main features of the MD data, but at a much better signal to noise ratio.

With this end in mind, we assume that the dynamics of the system that produces the time series $\vec{v}(t)$ can be expressed by the Langevin equation

$$\dot{\vec{v}}(t) = \vec{f}(\vec{v}(t)) + \vec{\eta}(t). \tag{1}$$

Here $\vec{f}$ describes the deterministic part of the dynamics, while $\vec{\eta}$ denotes a stochastic driving term which represents the fluctuations of all degrees of freedom we want to ignore, including, e.g., high-frequency bond oscillations, the motion of the solvent, and the realization of the external heat bath. To obtain a simple model for the deterministic part $\vec{f}$ of the dynamics, the following steps are taken. First, we restrict the analysis to the first d_{EL} PCs, thus disregarding all components accounting for simple Gaussian fluctuations. Since only the deterministic part is the subject of the dimensionality reduction [the noise term $\vec{\eta}(t)$ by definition explores all directions of phase space], we also neglect the stochastic driving in Eq. (1). This is realized by applying a simple noise reduction scheme, i.e., the Savitzky-Golay filter [17] to the resulting trajectory $\vec{v}(t) \in \mathbb{R}^{d_{\mathrm{EL}}}$. The third step is to construct a state space in which the trajectory $\vec{v}(t)$ shows a deterministic behavior. Since $\vec{v}(t)$ represents a projection of the original phase space (that explicitly includes the positions and momenta of all atoms of the system), in general we cannot expect that the dimension d_{EL} of the trajectory is sufficient for this purpose. To account for a possibly higher-dimensional phase space, we use an extension of the Takens embedding [18] and embed the first (and most important) PC until this component is reconstructed sufficiently well. Adding the remaining $d_{\mathrm{EL}} - 1$ components and using a time delay Δt, the resulting embedding vector at time step t_j reads

$$\begin{aligned}\vec{v}(t_j) &\equiv \vec{v}_j \\ &= (v_1(t_j), v_1(t_j - \Delta t), \ldots, v_1(t_j - (m \\ &\quad - 1)\Delta t), v_2(t_j), v_3(t_j), \ldots, v_{d_{\mathrm{EL}}}(t_j))^T,\end{aligned} \tag{2}$$

where m denotes the embedding dimension of the first PC, resulting in a dimension $d_{\mathrm{RS}} = d_{\mathrm{EL}} + m - 1$ for the reconstructed state space. For all systems considered, $m = 9$ and Δt from 0.2 ps (Ala$_3$) to 4 ps (Ala$_{10}$) were used.

Knowing the state space, we are now in a position to fit a deterministic nonlinear model to the data. Following Farmer and Sidorowich [19], we employ a locally linear model defined by the map

$$\vec{v}_{j+1} = \mathbf{A}_j \vec{v}_j + \vec{b}_j, \tag{3}$$

where $\mathbf{A}_j$ is a $d_{\mathrm{RS}} \times d_{\mathrm{RS}}$ matrix. Locally linear means that, given the vector $\vec{v}_j$ at time t_j, the subsequent vector $\vec{v}_{j+1}$ at time t_{j+1} is obtained in linear approximation, that is, the model parameters $\mathbf{A}_j$ and $\vec{b}_j$ are obtained by a least squares fit which only uses the spatial neighbors of $\vec{v}_j$ [16,20]. As a consequence, the model parameters need to be calculated for every time step of the model trajectory. To validate the model, we again consider the distributions and autocorrelation functions of the first two PCs of Ala$_3$ (Fig. 2) and Ala$_{10}$ (Fig. 3) and compare the modeled data to the results obtained from the MD simulations. Reproducing the time scales of the dynamics as well as the conformational distribution in almost all details, the model accounts nicely for the essential features of the MD data.

Let us now turn to the Lyapunov exponents λ_i, ($i = 1, \ldots, d_{\mathrm{RS}}$) of the peptide dynamics, which are calculated through the Jacobian matrix $\mathbf{A}_j$ of the map (3). For all systems considered, we found two positive exponents λ_1 and λ_2, which quantify the chaoticity of the dynamics in phase space. We first consider the Kolmogorov-Sinai entropy h_{KS}, which is given by the sum of the positive Lyapunov exponents [16]. Its reciprocal value $\tau_{\mathrm{KS}} = 1/h_{\mathrm{KS}}$ is an estimate for the time span the evolution of the trajectory can be forecasted. As shown in Table I, this picosecond time scale increases with system size, thus indicating that the structural dynamics of the larger peptides is less chaotic than the dynamics exhibited by the smaller systems.

To estimate the effective dimension d_{KY} from the Lyapunov exponents, we employ the Kaplan-Yorke conjecture [3]

$$d_{\mathrm{KY}} = k + \frac{1}{|\lambda_{k+1}|} \sum_{i=1}^{k} \lambda_i, \tag{4}$$

where k is the number of exponents such that (if they are ordered decreasingly) the sum of the first k exponents is still positive or zero, whereas the sum of the first $k+1$ exponents is already negative. Loosely speaking, the defi-

TABLE I. Comparison of d_{EL}, the dimension of the energy landscape, and d_{KY}, the effective dimension of the dynamics, as obtained for various alanine peptides. Also shown are n_{HB}, the average number of α_R-type intramolecular hydrogen bonds, and τ_{KS}, the reciprocal value of the Kolmogorov-Sinai entropy.

	Ala$_3$	Ala$_5$	Ala$_7$	Ala$_{10}$
d_{EL}	2	3	6	8
d_{KY}	5.0	4.7	4.9	3.3
n_{HB}	⋯	0.03	0.6	2.4
τ_{KS} [ps]	3.8	3.7	5.9	8.0

nition of the leading term k in d_{KY} assures that the phase-space expanding ($\lambda_i > 0$) directions just counterbalance the phase-space contracting ($\lambda_i < 0$) directions, thus warranting an overall invariant phase-space volume. Table I lists the resulting values of the effective dimension d_{KY} obtained for Ala_3 through Ala_{10}. Ranging from ≈ 3 to 5, the dimensions appear to be quite small, considering that it accounts for the motion of thousands of atoms. Most intriguing, though, is the fact that the effective dimension decreases with system size, from 5 for Ala_3 to 3.3 for Ala_{10}. This is in striking contrast to the behavior of the energy landscape dimension d_{EL} which, as expected, increases with chain length.

To explain this finding, detailed analyses of the all-atom MD trajectories were performed, which revealed that the effect is caused by intramolecular interactions that stabilize the secondary structure of the peptide. Most importantly, this is achieved by intramolecular hydrogen bonds connecting the nth and $(n + 4)$th residues of the amino acid chain, thus stabilizing the α_R helix structure (see Fig. 1). As shown in Table I, the average number of these hydrogen bonds increases significantly, once the number of possible α_R-type bonds reaches three for Ala_7. Remarkably, the formation of stabilizing hydrogen bonds seems to significantly reduce the effective dimension, although these bonds are not stable but formed and broken on a nanosecond time scale.

It is interesting to note that this decrease of the effective dimension is not observed for the energy landscape dimension d_{EL}. Apparently, this is because the latter quantity is defined in the linear framework of PC analysis theory, whereas the effective dimension d_{KY} is obtained from a nonlinear description of the dynamics. In a similar vain, other nonlinear methods for the analysis of biomolecular dynamics have been proposed that are sensitive to nonlinear correlations and therefore may reduce the dimensionality of the problem [21,22]. The effect of nonlinear dimensionality reduction is supposedly even more important for the folding of larger peptides and proteins, which exploit a variety of stabilizing interactions and exhibit significant cooperativity [23].

While the work presented here is only a first step towards a nonlinear analysis of MD data, it may open ways to address the larger problem of describing folding processes. For example, we wish to study if the folding of various structural motifs such as α_R helices and β sheets results in distinguishable properties of the corresponding dynamical model. Another next step is to go beyond the locally linear ansatz and construct analytical models of the dynamics. Such analytical models would contain a set of parameters which presumably depend on, e.g., experimental conditions, amino-acid sequence, and folding motifs. The study of this parameter dependence could then shed some light on the still elusive mechanism of folding.

We thank Alexandra Villa for numerous inspiring and helpful discussions. The work presented has been supported by the Frankfurt Center for Scientific Computing, the Fonds der Chemischen Industrie, and the Deutsche Forschungs-
gemeinschaft.

*Electronic address: hegger@theochem.uni-frankfurt.de

[1] J. D. Farmer, E. Ott, and J. A. Yorke, Physica (Amsterdam) **7D**, 153 (1983).

[2] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).

[3] P. Frederickson, J. L. Kaplan, E. D. Yorke, and J. A. Yorke, J. Diff. Equ. **49**, 185 (1983).

[4] D. Frenkel and B. Smit, *Understanding Molecular Simulations* (Academic, San Diego, 2002).

[5] J. N. Onuchic, Z. L. Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[6] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, U.K., 2003).

[7] T. Ichiye and M. Karplus, Proteins **11**, 205 (1991).

[8] A. E. Garcia, Phys. Rev. Lett. **68**, 2696 (1992).

[9] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins **17**, 412 (1993).

[10] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comput. Chem. **26**, 1701 (2005).

[11] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, and I. G. Tironi, *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (, Zürich, 1996).

[12] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans, in *Intermolecular Forces*, edited by B. Pullman (D. Reidel Publishing Company, Dordrecht, 1981), pp. 331–342.

[13] Y. Mu, P. H. Nguyen, and G. Stock, Proteins **58**, 45 (2005).

[14] Y. Mu and G. Stock, J. Phys. Chem. B **106**, 5294 (2002).

[15] V. Villani, J. Mol. Struct. **621**, 127 (2003).

[16] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge Univ. Press, Cambridge, U.K., 1997).

[17] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipies* (Cambridge University Press, Cambridge, 1992), 2nd ed..

[18] F. Takens, in *Dynamical Systems and Turbulence (Warwick 1980)*, edited by D. A. Rand and L.-S. Young, Lecture Notes in Mathematics Vol. 898 (Springer-Verlag, Berlin, Heidelberg, 1980), pp. 366–381.

[19] J. D. Farmer and J. J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).

[20] R. Hegger, H. Kantz, and T. Schreiber, Chaos **9**, 413 (1999).

[21] O. F. Lange and H. Grubmüller, Proteins **62**, 1053 (2006).

[22] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **103**, 9885 (2006).

[23] K. A. Dill, Biochemistry **29**, 7133 (1990).