RESEARCH ARTICLE

Quantifying the Complexity of Chaos in Multibasin Multidimensional Dynamics of Molecular Systems

DMITRY NERUKH, GEORGE KARVOUNIS, AND ROBERT C. GLEN
Unilever Centre for Molecular Informatics, Department of Chemistry, Cambridge University, Cambridge CB2 1EW, UK

Received February 2, 2004; revised September 22, 2004; accepted September 24, 2004

The simulated classical dynamics of a small molecule exhibiting self-organizing behavior via a fast transition between two states is analyzed by calculation of the statistical complexity of the system. It is shown that the complexity of molecular descriptors such as atom coordinates and dihedral angles have different values before and after the transition. This provides a new tool to identify metastable states during molecular self-organization. The highly concerted collective motion of the molecule is revealed. Low-dimensional subspaces dynamics is found sensitive to the processes in the whole, high-dimensional phase space of the system. © *2004 Wiley Periodicals, Inc. Complexity 10: 40–46, 2004*

Key Words: molecular self-organization; statistical complexity; molecular dynamics; chaos

INTRODUCTION

Molecular systems generally have complicated dynamics. Many-body and highly nonlinear interactions make most of the dynamics chaotic, which is especially true for large systems such as biomolecules. Stochastic behavior of the molecular motions is assumed in statistical mechanics; however, it is not random. It is chaotic and is generated by deterministic laws of motion. The nonrandom behavior is crucial when molecular self-organization is considered, e.g., in protein folding.

Correspondence to: Dmitry Nerukh, E-mail: dn232@cam.ac.uk

A very high number of degrees of freedom and/or complex interactions make the potential energy surface "rugged," that is, covered by many local minima with similar energies. This leads to a situation in which significant changes in the molecular configuration become "rare events." It can take a long time for a molecule to cross an energy barrier and find the next local minimum. Often the locations of some minima (intermediate conformations) are known and the problem posed is how to find the path in multidimensional space to reach one state from the other.

The latter problem attracts much attention because it holds the key, e.g., to understanding chemical reaction mechanisms. Finding the reaction path is a substantial challenge because often it cannot be probed directly by exper-

DOI 10.1002/cplx.20055

iment. The "rareness" of the event and ultimately the complexity of the dynamical system makes it difficult to investigate theoretically or simulate numerically. Attempts to rationalize the mechanisms of these "rare events" and find the reaction paths has suggested that the dynamics of the metastable states between the transitions is significantly chaotic, whereas at the moment of the transition it becomes semi-chaotic or quasi-regular, i.e., the system can maintain approximate constants of motion and possess fully deterministic dynamics [1]. An important question still remains: how to identify the stable and metastable states in between these transitions [2]. Where molecular fluctuations involve only a few degrees of freedom this can be done empirically or by using methods such as principal components analysis. However, when many degrees of freedom are interacting nonlinearly, recognizing the dynamic patterns of stable states can be difficult.

This is because chaos itself contains patterns in time. Processes having the same randomness can contain temporal structures of different degrees of complexity. Detecting chaos can be problematic. The widely used Lyapunov exponent is a necessary but not sufficient characteristic to identify chaos and has been reported to fail to detect chaos [3]. A more detailed exploration of the inner structure of chaos using, e.g., Komogorov's algorithmic complexity suggests alternative methods of detecting chaos [3].

The inner structure in a chaotic signal is reflected in the way the system explores its phase space. More structure means more repetitions of some kind in the trajectories of the phase space. This implies that not all regions of the space are uniformly covered. The evidence for this is indeed found for protein dynamics, which has been shown to be suppressed and cover less configuration space than normal Brownian (completely stochastic) process on a short time scale [4].

Numerous approaches to reduce dimensionality by extracting only the most important degrees of freedom have encountered significant problems and suggest that it is difficult to identify the "reaction coordinate" in this simple way [5]. Moreover, it is often necessary to introduce additional coordinates to capture the event studied. For example, the simulation of the alanine dipeptide in explicit water demonstrated the importance of the angle θ and coordinates of the water molecules in addition to the traditional dihedrals φ and ψ to identify the reaction coordinates for the process of transition to a helical turn [6]. Similarly, in addition to the fraction of native contacts Q, usually used as a reaction coordinate for protein folding, other coordinates are necessary to obtain an optimal dividing surface [7]. This and other investigations lead to the conclusion that generally, the processes in complex molecular systems can only be adequately described as "collective" motion, not by the dynamics of individual atoms or even groups of atoms. The many-body character of the dynamics plays a decisive role, especially in the condensed phase when solvent molecules are included.

In summary, when investigating the transition processes in molecular systems (i) the dynamics of many atoms must be taken into account simultaneously; (ii) very high dimensional trajectories of the system need to be analyzed; (iii) the details (inner structure) of the chaotic system's dynamics should be studied; and (iv) the different characters of the fluctuations should be considered: chaotic behavior between the transitions separated by quasi-regular or regular dynamics at the moment of transition.

Taking these points into account suggests that the dynamics of a complex molecular system can be investigated from a new perspective. The analysis of the dynamical trajectory directly with an attempt to discover the patterns and regularities in it rather than investigating the average quantities as is done in statistical physics is a new approach. These patterns constitute the essence of complexity [8] and, to our view, provide the missing perspective on molecular systems: the investigation of dynamical complexity and its emergence in the system. This is especially important for self-organizing systems because complexity is intuitively connected to the emergence of new structures.

More specifically, we calculate the "statistical complexity" of various molecular dynamic parameters such as atomic trajectories, reciprocal orientations, and dihedral angles. Statistical complexity is a characteristic developed within the framework of "computational mechanics" [9–11]. We have shown that it can be used as a valuable tool for uncovering new information about molecular processes in the condensed phase [12,13].

Related to "computational mechanics" is "evolutionary dynamics" [14]. This considers the time evolution of a population as a series of long-lasting "epochs" when the overall fitness of the population remains approximately the same and rapid "innovations," when the system finds a way to a new higher level of fitness [14]. Besides biological evolution, this developing hypothesis describes a wide range of systems from ferromagnetic spin systems to genetic program algorithms [15]. Here the interplay between the competing forces of order and disorder leads to the emergence of randomness and structure during the chaotic dynamics. The hypothesis suggests that the system's phase space is partitioned into sub-basins (basins of attractions), where the system spends most of the time (epochs, periods of stasis), slowly exploring a sub-basin with almost unchanging values of "fitness" (e.g., energy). At some rare moments the system finds a "portal" to another sub-basin with a higher fitness. The system quickly transfers most of its population to this new sub-basin and continues to explore until the discovery of the next "portal" [14,16]. This view on evolving dynamics contrasts with more common approaches when evolution is considered to be a more or less steady ascent from one fitness mountain to another.

We believe that the concept can be applied to molecular systems in cases where the time transformations of the system are "rare events" leading to some sort of self-organization. We see the evidence for this behavior in protein folding in which the dynamics is shown to be an ensemble of nearly degenerate substates and transitions between them [17].

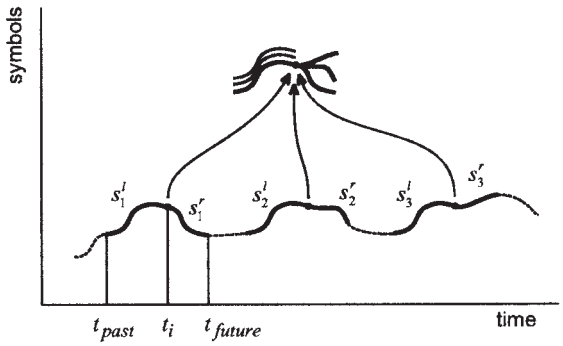
It must be emphasized that choosing the "fitness" function is a key point in the proposed research. As suggested by Crutchfield et al. [14] the focus should be on how the system stores and transforms information (rather than energy). The idea is that the intrinsic computational property (dynamic complexity) of the system is of primary importance. This suits our situation particularly well because the complexity is exactly the feature that discriminates one chaotic regime from another. Thus, dynamical complexity not only serves as a new valuable tool for tackling the extremely technically complex problem of analysing multidimensional dynamic signals but also plays an important conceptual role. It is suggested that the complexity of a dynamic system emerges (rises) when it discovers portals between the sub-basins in its phase space. These innovations are accompanied by changes in the architecture of information processing. Therefore, the analysis of the dynamic complexity of molecular system can shed a light on the details of the sub-basin– portal architecture in very high dimensional phase space.

The purpose of this work is to *study the chaotic motion of* a model *molecular system* within the framework of *computational mechanics*, thus investigating the *hypothesis of "epochal" evolution* punctuated by rapid innovations to new stable states. We have chosen a model that is sufficiently simple that method for analysis of complexity applied to molecular dynamics simulation can be successfully developed.

METHOD AND NUMERICAL SIMULATION

In the following symbolic dynamics is considered, i.e., the signal consists of discrete symbols assigned to discrete time steps. Let a set of symbols corresponding to each time step t_i form a sequence S. To calculate the statistical complexity [9–11] S is decomposed into a set of left s_i^l (past) of length l and right s_i^r (future) of length r halves joined together at time points t_i. Consider a particular left subsequence s_1^l and all left subsequences equivalent to it: s_2^l and s_3^l. Collect a set of all right subsequences following this unique left subsequence (Figure 1). Each right subsequence has its probability conditioned on the particular left one: $\Pr(s^r|s_i^l)$. The equivalence relation between any two left subsequences can now be defined. Two unique left subsequences s_i^l and s_j^l are equivalent if their right distributions are the same up to some tolerance value δ: $\Pr(s^r|s_i^l) = \Pr(s^r|s_j^l) + \delta$. A set of all equivalent left subsequences forms an "equivalence class." The equivalence classes have their own probabilities (A_i)

FIGURE 1



A schematic representation of the equivalence relations. The left ("past") subsequences s_1^l, s_2^l, and s_3^l (all symbols on the $[t_{past}, t_i]$ interval) are the same. They lead to a distribution of right ("futures") subsequences s_1^r, s_2^r, and s_3^r ($[t_i, t_{future}]$).

calculated from the probabilities of the constituent left subsequences. In all our calculations the tolerance of 0.001 was used (see [18] for details).

The importance of the notion of equivalence classes is that they represent the states of the system that define the dynamics at future moments—the "causal states" (here equal to "equivalence classes" with corresponding probabilities). The time evolution of the system can be viewed as traversing from one causal state to the other with a probability defined by $\Pr(s^r|s_i^l)$. The set of the causal states together with the transition probabilities constitute a so called "ϵ-machine." ϵ-machines represent the minimal computation necessary to reproduce the dynamics of the system [18].

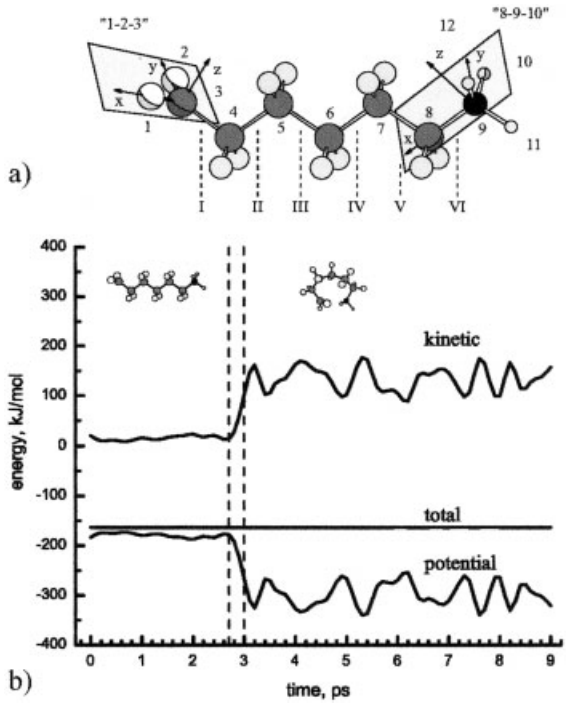
The statistical complexity is defined as the informational size of the ϵ-machine. The measure of this is the Shannon entropy of the causal states:

$$C \equiv -\sum_{A_i} \Pr(A_i)\log_2 \Pr(A_i),$$

where A_i are causal states. In contrast to Kolmogorov complexity this measure provides a zero complexity for *both* extremes: a constant signal and a purely random process. The maximum value of complexity lies somewhere in between these two limits.

The essence of statistical complexity is in the analysis of the symbolic dynamics based on a physical process. This means that the real signal is converted into a series of discrete symbols from a finite alphabet using appropriate partitioning of the phase space. The difficulties encountered and the technical details of the symbolization procedure used can be found in [12]. Omitting the details, the symbolization was done by dividing the phase space into the grid of

FIGURE 2



(a) Schematic view of the zwitterion. Carbon atoms are dark grey; hydrogen, light grey; oxygen, white; nitrogen, black. The numbers label the atoms of the model, the Roman numerals are the dihedral angles. (b) Energy is in kJ/mol of the system simulated. Vertical dashed lines show the beginning (2.7 ps) and end (3.0 ps) of the transition.

a specified coarseness (partitioning) and assigning the symbols by the cells where a data point falls.

The molecular model we investigate was chosen to be simple yet complex enough to represent the basic features of a multiatomic molecular system with nonequilibrium, self-organizing behavior. We simulate the classical dynamics of a zwitterion with charged oxygen and nitrogen atoms in a vacuum (Figure 2). The Gromos-96 [21] force field and LINCS [22] algorithm were used (the united atom model, considering the CH_2 groups as one particle, with all bond lengths constraint). The GROMACS Molecular Dynamics program [19,20] was used for all simulations. The system was initially prepared in the extended configuration, and then its energy was minimized and an MD equilibration run took place for 2000 steps of 1 fs. After this the data were collected from the MD simulation run for 8000 steps of 1 fs. In the classical model adopted, there are no proton transfer reactions allowed. In a real system in the gas phase, a proton transfer may be observed from the protonated amine to the carboxylic acid.

This model Hamiltonian system possesses 19 degrees of freedom with highly nonlinear interactions that with high probability implies that its dynamics is chaotic. Indeed, for molecular systems it has been shown that molecular systems exhibit chaotic behavior (have positive Lyapunov exponents) both for large bio-molecules [23,24] and simple three-atomic molecular model [25]. In addition, as it will be shown later, this system, despite of its simplicity, also has distinct "basins of attraction" and features of the "rare events" dynamics.

For the analysis we collected the system's dynamical parameters in three ways: (1) three-dimensional (3D) trajectories of each of the 12 atoms of the molecule (Figure 2; called "global" further in the text); (2) using the same 3D trajectories but in a "local" coordinate systems attached to each end of the molecule (Figure 2; called "1-2-3" and "8-9-10"); and (3) six (1D) dihedral angles (Figure 2).

RESULTS AND DISCUSSION

Visualization of the zwitterion motion shows initial fluctuations in an "extended" conformation. The molecular chain is more or less stretched along a straight line. This is followed by rapid collapse into a "folded" state, when the charged ends are next to each other and the molecule forms a ring-like structure (Figure 2). By plotting the system's energies, the evidence of a clear structural transformation can be seen in Figure 2.

It should be stressed that we did not use any temperature-controlling mechanisms in order to keep the total energy constant. This leads to a substantial increase of kinetic energy after the folding that compensates the potential energy drop. However, the important point is that the system remains on the same energy level and irreversible character of the system's dynamics presents a good example of a relatively simple dynamic system with nonergodic behavior.

In our model the moment of transition is clearly defined by both the geometry transformation and energy changes, and we can proceed with testing our hypothesis and check if the complexity will indicate the same moment of transition. We assign the time from 2.7 to 3.0 ps as a period of transition and the stages before and after this period as the times when the system is in different "phase space basins."

In the case of a continuous trajectory, i.e., when the size of the alphabet goes to infinity, existence of probabilities $\Pr(s^r|s_i^l)$ assumes that at a particular time t_i the system can follow different trajectories starting from a single point in the phase space. That is obviously not the case for the Hamiltonian deterministic systems where the trajectory is uniquely defined by the initial conditions. Therefore, the causal states will degenerate into single pairs of past and futures and the probabilities of the causal states will reflect the probabilities of the system to be found in each phase space point over a period of time under consideration. However, this is not the case when we consider a subspace, rather than a whole phase space. Moreover, from the chemical point of view, considering different subsystems of the

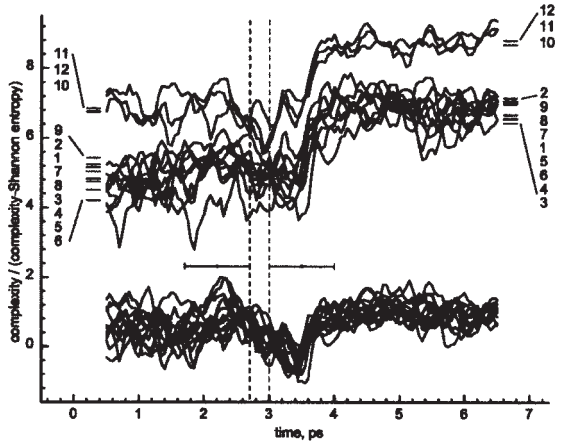
whole system can be even more informative because often in molecular systems only relatively small fraction of atoms exhibits a desired behavior. For example, for a protein in water only the protein molecule shows folding, whereas the majority of the degrees of freedom, although being necessary for the folding to occur, —the water molecules—wander in the phase space without forming persistent structures.

The above is true only for the limit of infinitely fine partitioning, i.e., for continuous signals. For a finite size of partitioning grid, however, each particular cell in the phase space may lead to different trajectories because of the range of possible initial conditions within the cell. Therefore, the partitioning of the phase space to symbolize the continuous signal should be optimal in a sense that there should not be too many partitions leading to a situation described above and, at the same time, should be enough partitions to extract the relevant information from the continuous signal. We found empirically that a reasonable alphabet was 50 partitions in each of the three dimensions for "global" trajectories and 25 partitions for "local" trajectories and dihedrals.

The 3D trajectories of the atoms (not reproduced here) show two regions of chaotic behavior with seemingly random character. "1-2-3" and "8-9-10" local trajectories reflect the dynamics of the atoms excluding the translations and rotations of the molecule as a whole. The difference between the two is in their ability to provide dynamic data for different ends of the molecule: the closer an atom to the origin, the smaller the absolute value of the coordinate, and, consequently, less information is transferred into the symbolic sequence. The dihedrals show typical "rare events" in their temporal behavior (not shown here)—they tend to fluctuate around particular values with occasional jumps from state to state. To calculate the complexities we used subsequences of 1 ps in length and calculated the statistical complexity on those intervals using the procedure described above. The resulting complexity value was plotted as a point in the middle of the interval, i.e., at 0.5 ps. The procedure was repeated to cover the whole simulation time. The length of the interval (1 ps) used to calculate each complexity value is the reason of the appearance of a delay in complexity change after the transition took place.

The complexity of the "global" trajectories together with the complexity–Shannon entropy difference are shown in Figure 3 (Shannon entropy was calculated using the probabilities of the symbols obtained from the same symbolization procedure as for the complexity calculation). The first and foremost result is that the complexity demonstrates a distinctive increase after the transition point. Most interestingly, all 12 atoms exhibit similar changes in complexity. This suggests that the dynamics in all 19 dimensions for the system (we do not include velocities at the moment) is important. Amazingly, individual 3D components are sensitive to the whole 19-dimensional process. This suggests that complexity can be a powerful tool in identifying the states of a very high dimensional system by using trajectories in a very small dimensional subspace.

FIGURE 3



Statistical complexity of the atoms ("global" trajectories; see text) of the model system investigated. The numbers on the left and right with corresponding short horizontal lines depict the mean values of complexity for each atom (see Figure 2a) before and after the transition. The upper group of curves (above 2) are statistical complexities; the lower group are the differencies between the complexities and Shannon entropies. The dashed lines indicate the transition interval (see Figure 2b). The horizontal error bars show the intervals used for calculating each point of complexity.

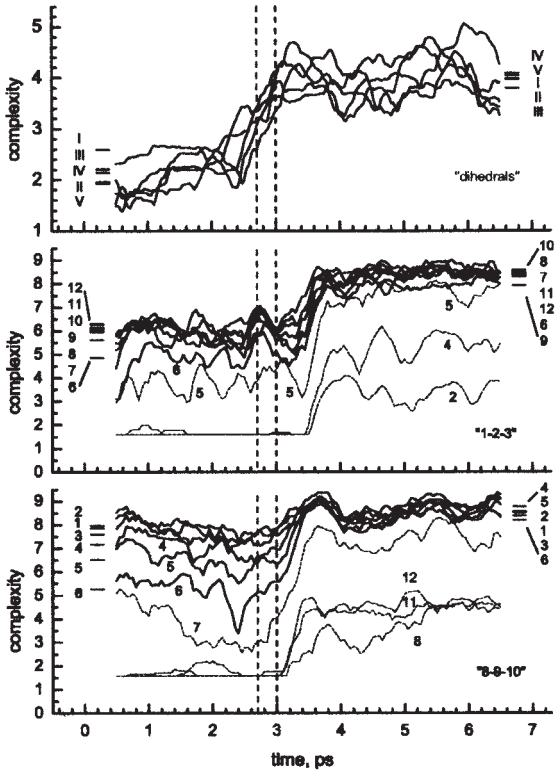
It can be seen from Figure 3, that statistical complexity provides information that is different from Shannon entropy, demonstrating that in addition to the change in the randomness of the motion, the information of temporal history plays an important role. Statistical complexity carries information about the emergent patterns of the process, so, e.g., a completely random sequence has zero complexity (having the maximum Shannon entropy).

The hydrogen atoms 10, 11, and 12 show slightly higher complexity. This could be because they perform faster rotations around the CN bond which, apparently, introduces more dynamic features to their trajectories.

After the collapse, the complexities seem to be closer to each other compared to the regimes before the transition. This implies that in the "folded" state the dynamics of the system is more concerted and all atoms are involved in the same highly interdependent motion.

The latter effect is even more pronounced for the "local" trajectories (Figure 4). The same qualitative behavior is present (the absolute value of complexity strongly depends on the number of partitions used) as is seen in the "global" trajectories. The "1-2-3" trajectories reproduce a distinctive peak just before 3 ps, also present at some of the "global"

FIGURE 4



Same as Figure 3 but for the "local" and "dihedrals" trajectories. The dotted curves are for the trajectories that are close to the coordinate system origin (see text for details).

atoms' complexities. The less pronounced changes in the values of complexity of the "8-9-10" trajectories is a result of the NH_3 group rotations that appear to possess somewhat different dynamics when compared to the rest of the molecule.

Finally, the biggest change in complexity is found for the dihedral angles (Figure 4). For the reasons discussed in the previous paragraph, we omitted the dihedral angle number VI, which corresponds to the rotations of the NH_3 group. Obviously, for this sort of molecular model, the dihedrals reflect most of the features of molecular configuration, which is clearly reflected in their behavior. This is beneficial since the dihedral angles are one-dimensional, which allows the accumulation of more information from the continuous trajectories during the symbolization. Again it should be stressed that all the dihedrals show qualitatively the same behavior at the moment of transition. This supports the previous conclusion that the dynamics and hence the complexity is highly concerted involving all atoms with a strong interdependence.

It should be mentioned that the system undergoes a substantial increase in temperature defined in the usual statistical mechanical way. Therefore, the observed changes in complexity can be at least partially attributed to the temperature rise. To test in what extent the temperature affects the complexity additional investigation is required that will involve the simulation with temperature coupling. However, this will change the system completely lifting the Hamiltonian systems restrictions and introducing the dissipative forces in it. This, in turn, fundamentally changes the dynamic picture and the analysis will require significantly different approach. We will explore this line of research in our subsequent publications.

CONCLUSIONS

The main result of the present study is in presenting evidence that substantiates the sub-basins—portal architecture of a molecular system's dynamics, at least for the model used. The increase in dynamical complexity when the system undergoes a transformation leading to a more complex structure is demonstrated. The highly concerted collective motion of the system is revealed and, the 3D subspaces, or even 1D cuts if an appropriate coordinate transformation is applied are sensitive to the processes taking place in the very high-dimensional phase space. The statistical complexity provides a valuable new tool for discriminating between the regions of the dynamical system when they are at different metastable states during the transition over time.

The results presented give us a promising new direction for the analysis of more complex molecular systems, e.g., biomolecules in water. We anticipate the situation where the complexity of atoms in different locations of a large molecular system during folding have similar complex behavior. This would provide a fundamentally new way of understanding the transition processes in the system. When a system has reached its most structured state, (e.g., when a protein has folded), we hypothize that the dynamical behavior of many of its atoms involved in secondary and tertiary structural features would have similar complexity behavior.

ACKNOWLEDGEMENT

This work is supported by the Isaac Newton Trust and Unilever.

REFERENCES

1. Komatsuzaki, T.; Berry, R.S. Adv Chem Phys 2002, 123, 79.
2. Bolhuis, P.G.; Chandler, D.; Dellago, C.; Geissler, P.L. Annu Rev Phys Chem 2002, 53, 291.
3. Crisanti, A.; Falcioni, M.; Mantica, G.; Vulpiani, A. Phys Rev 1994, E50, 1959.
4. Garcia, A.E.; Hummer, G. Proteins 1999, 36, 175.
5. Onuchic, J.N.; Nymeyer, H.; Garcia, A.E.; Chaine, J.; Socci, N.D. Adv Protein Chem 2000, 53, 87.

6. Bolhuis, P.G.; Dellago, C.; Chandler, D. Proc Natl Acad Sci USA 2000, 97, 5877.
7. Plotkin, S.S.; Wolynes, P.G. Phys Rev Lett 1998, 80, 5015.
8. Zurek, W., Ed. Entropy, Complexity, and Physics of Information, SFI Studies in the Sciences of Complexity, Vol. VIII; Addison-Wesley: Reading, MA, 1990.
9. Crutchfield, J.P.; Young, K. Phys Rev Lett 1989, 63, 105.
10. Crutchfield, J.P.; Young, K. Entropy, Complexity, and Physics of Information, SFI Studies in the Sciences of Complexity, Vol. VIII; Zurek, W., Ed.; Addison-Wesley: Reading, MA, 1990.
11. Crutchfield, J.P. Physica D 1994, 75, 11.
12. Nerukh, D.; Karvounis, G.; Glen, R.C. J Chem Phys 2002, 117, 9611.
13. Nerukh, D.; Karvounis, G.; Glen, R.C. J Chem Phys 2002, 117, 9618.
14. Crutchfield, J.P. Evolutionary Dynamics. Epochal Evolution. Exploring the Interplay of Selection, Accident, Neutrality, and Function; edited by Crutchfield, J.P.; Schuster, P., Eds.; Oxford University Press: New York, 2002.
15. Crutchfield, J.P.; Schuster, P., Eds. Evolutionary Dynamics. Exploring the Interplay of Selection, Accident, Neutrality, and Function; Oxford University Press: New York, 2002.
16. Crutchfield, J.P.; van Nimwegen, E. Evolution as Computation; Landweber. L.; Winfree, E., Eds.; Springer-Verlag: New York, 2001.
17. García, A.E. Phys Rev Lett 1992, 68, 2696.
18. Shalizi, R.C. Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata, PhD thesis, University of Wisconsin at Madison, 2001.
19. Berendsen, H.J.C.; van der Spoel, D.; van Drunen, R. Comp Phys Comm 1995, 91, 43.
20. Lindahl, E.; Hess, B.; van der Spoel, D. J Mol Mod 2001, 7, 306.
21. van Gunsteren, W.F.; Billeter, S.R.; Eising, A.A.; Hunenberger, P.H.; Kruger, P.; Mark, A.E.; Scott, W.R.P.; Tironi, I.G. Biomolecular Simulation: The GROMOS96 Manual and User Guide; Hochschulverlag AG an der ETH Zurich, Zurich, 1996.
22. Hess, B.; Bekker, H.; Berendsen, H.J.C.; Fraaije, J.G.E.M. Lines: A linear constraint solver for molecular simulations. J Comp Chem 1997, 18, 1463.
23. Braxenthaler, M.; Unger, R.; Auerbach, D.; Given, J.A.; Moult, J. Chaos in biomolecular dynamics. Proteins Struct Funct Genet 1997, 29, 417.
24. Zhou, H.-b. Chaos in protein dynamics. J Phys Chem 1996, 100, 8101.
25. Rage, T.; Neumaier, A.; Schlier, C. Rigorous verification of chaos in a molecular model. Phys Rev E 1994, 50, 2682.