Computational Mechanics reveals nanosecond time correlations in molecular dynamics of liquid systems

Dmitry Nerukh[a]

[a] Unilever Centre for Molecular Informatics, Department of Chemistry, Cambridge University, Cambridge CB2 1EW, UK

Statistical Complexity, a measure introduced in Computational Mechanics has been applied to MD simulated liquid water and other molecular systems. It has been found that Statistical Complexity does not converge in these systems but grows logarithmically without a limit. The coefficient of the growth has been introduced as a new molecular parameter which is invariant for a given liquid system. Using this new parameter extremely long time correlations in the system undetectable by traditional methods are elucidated. The existence of hundreds of picosecond and even nanosecond long correlations in bulk water has been demonstrated.

1. Introduction

It is commonly believed that time correlations in liquids under normal conditions do not exceed several picoseconds. At sufficiently large times the dynamics of molecules can be described as ordinary diffusion, that is a purely stochastic process indistinguishable from noise. Indeed, the standard correlation function (calculated both from experimental and simulated data) analysis [1] shows that all velocity correlations in water vanish after ≈0.2 ps for hydrogen and ≈0.7 ps for oxygen (Fig. 1).

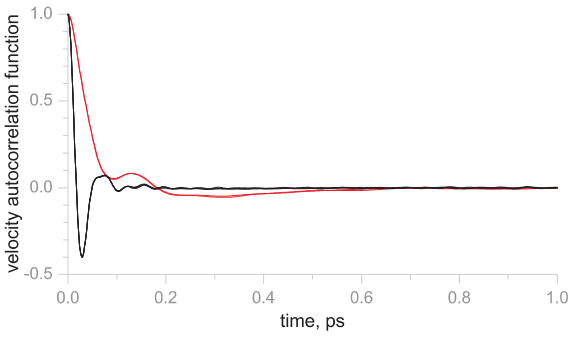

Figure 1. Velocity autocorrelation function for oxygen (red) and hydrogen atoms of two water molecules calculated as time average over 2000 ps. The curves for the atoms of the same type are practically indistinguishable.

Time correlations in atom coordinates are commonly quantified using the diffusion coefficient D defined through the mean square displacement of an atom $\langle x^2(t)\rangle \propto t^D$, where x is an atom's coordinate. For long enough times, D is equal to unity and the time changes of x are completely described as a random process that is zero correlation process. For water, the times when the correlations vanish and D becomes equal to unity are also of the order of several picoseconds.

Recent analysis pushes this boundary towards tens of picoseconds [2]. The authors investigated the moments (higher than two) of the displacement of atoms in MD simulated water and argon. They have demonstrated that for these times the moments significantly deviate from the behaviour predicted by the diffusion theory. It is shown that the process can be described by the continuous time random walk (in contrast to the simple random walk for normal diffusion) model [3] implying non-Markovian character of the dynamics. However, at the times exceeding several tens of picoseconds the system produces a simple Markov chain.

In this paper, we demonstrate the existence of dynamical time correlations in liquids at the time scale of hundreds of picoseconds and even nanoseconds. This is done by applying a special

statistical measure to the molecular signals, Computational Mechanics [4–6].

2. Molecular system

We have simulated bulk water (periodic boundary conditions) consisting of 392 SPC (Simple Point Charge) [7] molecules using the GROMACS molecular dynamics [8] package. The temperature of the system was kept constant at 300K using Berendsen [9] thermostat. Various number of molecules, water models, thermostat types and their parameters were investigated to check the consistency of the results. A sufficient equilibration was performed before collecting data for analysis. The velocity of one of the hydrogens was used as a signal for the analysis.

3. The method

3.1. Test on long time correlations

Before describing the details of the statistical measure that we used in the analysis, let us outline a general and simple procedure that can confirm the existence of dynamical correlations in a molecular system at times τ (we will refer to this procedure as "τ-test").

Suppose the molecular system at equilibrium generates a signal of the total length T significantly longer than expected correlation times τ: $T = n\tau$, where $n \gg 1$ (we will designate it as "original" signal). Let the signal be, for concreteness, the coordinate of one of the atoms of the system. A statistical measure applied to the signal produces a value M (for example, the diffusion constant D). It is important that the same value M is obtained for any realisation of the trajectory of length T, that is it is independent of the initial condition. Imagine also an ensemble of n realisations of the same system each starting from randomly chosen initial conditions. A test signal can be constructed from the ensemble such that it consists of τ-long pieces of the signals one from each realisation and the total length of the test signal equals to T (Fig. 2). The same statistical measure can now be calculated from this test signal and it would result in a value M' (D' for our example).

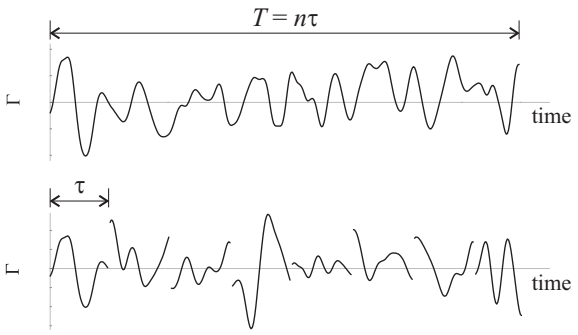

Figure 2. Schematic representation of two molecular trajectories used in the test revealing correlations at time τ (τ-test, see text). Γ represents the whole phase space of the molecular system. The top trajectory generates the "original" signal, the bottom one generates the "test" signal

Two situations are now possible. If the system does not exhibit correlations at time τ then the original and the test signals are statistically the same and M' should be equal to M ($D = D'$). This is because in the absence of correlations changing the molecular positions and velocities from time t to time $t + \tau$ is equivalent to randomly choosing new values of the coordinates and the velocities. On the other hand, any differences in the values of M and M' indicate that the original and the test signals are statistically different. Since the statistics on each piece separately are equal to each other (they are independent of the initial conditions), the differences between M and M' are the result of the changes introduced by the random shifts of the coordinates and velocities after the evolution of the system over the period τ. In other words, statistical correlations in the system at the time scale τ are present. This test is realisable in MD simulations since trajectories starting with specified initial conditions can easily be generated.

3.2. Statistical Complexity

The velocity autocorrelation function $f_{\mathbf{v}}(\tau) \equiv \frac{1}{T}\sum_t^T \mathbf{v}_t \cdot \mathbf{v}_{t+\tau}$ is a two point, linear statistical measure. Computational Mechanics [4–6] is

conceptually different because it operates on histories of $\mathbf{v}$. It analyses the histories ("pasts") $\{\ldots \mathbf{v}_{t-2}\mathbf{v}_{t-1}\mathbf{v}_t\}$ by grouping them into classes, called "causal states" ϵ_j, if the histories are followed by the same future $\{\mathbf{v}_{t+1}\mathbf{v}_{t+2}\ldots\}$ (probabilistically). Thus, the dynamics of the system is described by the probabilistic transitions between the causal states. Importantly, the statistic generated this way is a *unique minimal sufficient* statistic. This means that it is the most compact complete statistical description of the data, and it is also unique. The collection of the causal states together with the transition probabilities between them is called an "ϵ-machine". The rigorous definition of the ϵ-machine and its essential mathematical properties are provided in Appendix A.1.

The *Statistical Complexity*, C_μ, is the informational measure of the size of the ϵ-machine and quantifies the amount of information about the past of the system that is needed to predict its future dynamics: $C_\mu = H[P(\epsilon_j)]$, where H is the Shannon entropy of the distribution of a random variable X, $H[P(X)] \equiv -\sum_X P(X)\log_2 P(X)$, and $P(\epsilon_j)$ is the causal state probability (see Appendix A.1). ϵ-machines can be reconstructed from observed data using the `CSSR` algorithm described and implemented in [10].

Computational Mechanics analyses symbolic dynamics. In applying Computational Mechanics to molecular systems, a correct procedure of converting continuous molecular signals into a discrete symbolic sequences from a finite size alphabet has to be developed. The procedure we used is described in Appendix A.2.

4. Results and discussion

We first investigated the behaviour of ϵ-machine as a function of the length of the molecular signal, that is the simulation time. Much to our surprise, we have found that the causal states structure, the ϵ-machine, *never converges* at least at the lengths T of feasible MD simulations [11]. Instead, new causal states appear as more data are added during the simulation. This means that the system produces statistically different futures for the same pasts at all times observed in the simulation. Consequently, the value of Statistical Complexity grows with T. For water the dependence has a clear logarithmic character after ≈ 0.4ns, Fig. 3 (the initial high values of C_μ are due to the effect of the lack of data at small T when most of the sequences seen by the algorithm are unique that results in a large number of spurious causal states).

Similar behaviour can be observed for, for example, the mean square displacement of atoms. It also diverges with time, that is it goes to infinity at inifintely long times. The coefficient of the divergence is the diffusion constant D. Similarly, we have introduced the coefficient of the growth of Statistical Complexity, h_Q (Fig. 3): $C_\mu = a + h_Q \log_2 T$.

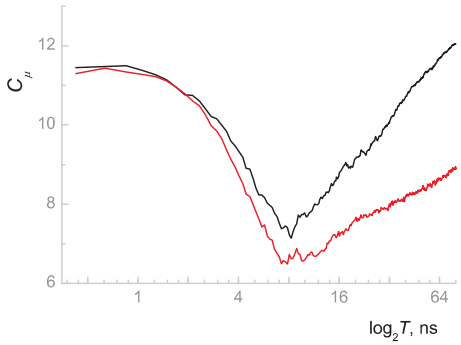

Figure 3. Statistical Complexity for the hydrogen velocity signal of bulk water. The original uninterrupted signal (red) and an ensemble consiting of 158ps long parts are shown

It should be stressed, that this behaviour is not an artefact of the procedure of the analysis, but rather an intrinsic property of the molecular dynamical system (see [11] for details). We have also verified that the value of h_Q does not depend on the details of the model of the liquid such as the number of molecules, water model types, parameters of the thermostat, etc (see [11] and Appendix A.3).

The phenomenon is observed not only in water but also in other molecular systems. For exam-

ple, for liquid argon $h_Q = 0.32$, while for a cluster of three water molecules in vacuum $h_Q = 0.90$. Therefore, h_Q seems to depend on the nature of the molecular system, that is on the system's inter-particle interactions. h_Q appears to be an invariant statistical characteristic of the liquid, similar to the diffusion coefficient. However, in contrast to the latter and the majority of the statistical descriptors of liquids, h_Q behaves fundamentally different in the τ-test described in section 3.1.

We have composed the test signal from the pieces of 158 ps long each and of the same total length as the original signal, 60 ns. While the autocorrelation function and the diffusion coefficient produce the same values for both signals, h_Q has significantly higher value for the composite signal (0.71 for the original signal and 1.21 for the test one). This means that the data in the pieces are statistically (in the Statistical Complexity sense) different between each other. We would like to stress again that each piece is significantly longer than the correlation times quantified by the traditional methods. Therefore, the test signal is indistinguishable from the original one from the point of view of the traditional analysis. However, hidden correlations do exist in the signal and can be discovered using the h_Q measure.

Thus, Computational Mechanics detects correlations in the molecular signals at the times of at least 158 ps. We have also tested longer pieces. h_Q remains higher than for the original signal even though the difference is smaller. This is because for longer pieces there are fewer statistically different parts of data and, consequently, it is statistically closer to the original signal.

In order to interpret the phenomenon from the point of view of molecular motion let us emphasize three observations concerning the phase space structure of the system.

First, the total simulation time is long enough for every molecule to cross the entire simulation box a few dozens of times. In other words, every degree of freedom of the system changes its value in the whole allowed range of possible values (the velocities oscillate from minimal to maximal values very quickly), that is the allowed phase space area is spanned from boundary to boundary.

Second, because of very high dimensionality of the system the phase space volume is extremely large. Even taking into account the fact that not the whole volume is explored but only the energetically allowed areas, the area accessible for the molecular trajectory is still very large. Therefore, the molecular trajectory never returns to itself, that is every time the trajectory crosses the phase space box it almost certainly takes a new route. This also follows from the chaotic nature of the system. However, different parts of the trajectory should be statistically close to each other because they are grouped together into a small number of causal states (the number of causal states is ≈ 400 for the longest trajectory of 60 ps that contains approximately two million histories).

Third, we have calculated the values of h_Q for different signals generated by the system: the velocities of the oxygen and hydrogen atoms and the instantaneous temperature in bulk water. The resulting values are the same within numerical errors. This shows that the phenomenon manifests itself in different degrees of freedom of the system and even in their combination (the temperature) [11]. Therefore, it is reasonable to assume that it reflects the behaviour of the full dimensional trajectory rather than the properties of individual signals produced by the system.

Taking into account these considerations we suggest the following microscopic picture. The molecular trajectory moves along a "network" of allowed "channels" in the phase space (Fig. 4). The "width" of the channels is relatively small and, consequently, the explored phase space volume of the system is also small, at least compared to the whole energetically allowed area of the phase space. That is why the rate of appearance of statistically new histories is significantly slower than it would have been in the case of completely independent histories from a trajectory uniformly covering the phase space. In the latter case every new history would form a new causal state and the size of the ϵ-machine would be equal to the total number of histories in the signal.

From this point of view h_Q quantifies the rate with which the trajectory explores the phase space restricted by the "network", and the rate

is much less than it could have been in the absence of any preferable routes in the phase space.

This picture explains why the τ-test signal produces larger value of h_Q. Since the relative volume of the network is much smaller than the whole allowed area of the phase space then initial conditions for the next piece chosen randomly fall with high probability outside the network of the current piece. In other words, for each piece in the test signal there exists its own network. When many networks are superimposed in one signal, the total area covered by the trajectory becomes larger than for the original signal (Fig. 4). Therefore, the phase space is explored faster which, in turn, results in a higher value of h_Q.

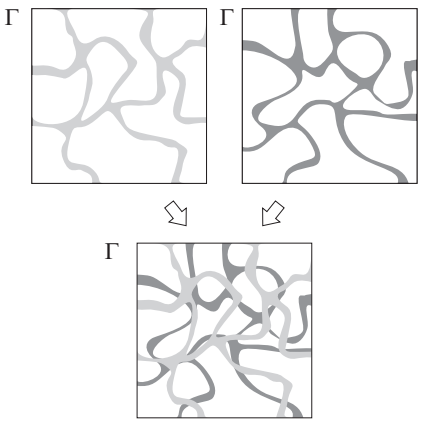

Figure 4. Schematic representation of the molecular phase space Γ and "network" formed by the trajectories for each uninterrupted part of the trajectory (see text). Upper two figures represent the networks for two individual parts while the lower figure demonstrates the larger phase space area of the combined network for the test signal phase space

5. Conclusions

Summarising, using Statistical Complexity, a measure on the histories of molecular signal, a parameter h_Q, is introduced. The parameter appears to be an invariant statistical characteristic of the liquid. It is fundamentally different from other commonly used statistical characteristics such as autocorrelation function or diffusion coefficient in that h_Q reveals hidden correlations in liquids on the time scale of hundreds of picoseconds and even nanoseconds, an order of magnitude longer than commonly accepted.

We conjecture that h_Q quantifies the rate with which the system's trajectory explores the phase space. In this way, it elucidates the structure in the full-dimensional phase space that is not visible to other methods. The structure appears to be a network in the phase space that the trajectory preferably follows. The area covered by this network (and by the trajectory evolving on it) is much smaller than the whole energetically allowed phase space area.

The network structure depends on the initial conditions. From the molecular point of view it means that it does matter where to start the trajectory. In the process of following the network structure the trajectory is confined within the network that introduces the very long time correlations. This is also a manifestation of non-ergodicity in the molecular system for which it is least expected to be found.

From the practical point of view the presented results can be important in cases where long lasting phenomena involving few molecules are under consideration. An obvious example is protein folding, where it becomes possible to experimentally analyse few (if not a single one) molecules over the time scales of nanoseconds. Other examples are from the field of various kinds of single molecular experiments, including single molecular spectroscopy [12]. Here the experimentally measured value is exactly of the kind described in this work: it is a signal generated by one unperturbed molecular process rather than an ensemble averaged quantity.

Acknowledgments. The work is supported by Unilever and the European Commission (EC Contract Number 012835 - EMBIO). DN thanks Professor Vladimir Ryabov for useful discussion on interpreting the results.

A. Appendix

A.1. Computational Mechanics

All past s_i^- and future s_i^+ halves of bi-infinite symbolic sequences centred at times i are considered. Two pasts s_1^- and s_2^- are defined equivalent if the conditional distributions over their futures $P(s^+|s_1^-)$ and $P(s^+|s_2^-)$ are equal. A *causal state* $\epsilon(s_i^-)$ is a set of all pasts equivalent to s_i^-: $\epsilon_i \equiv \epsilon(s_i^-) = \{\lambda : P(s^+|\lambda) = P(s^+|s_i^-)\}$. At a given moment the system is at one of the causal states, and moves to the next one with the probability given by the transition matrix $T_{ij} \equiv P(\epsilon_j|\epsilon_i)$. The transition matrix determines the asymptotic causal state probabilities as its left eigenvector $P(\epsilon_i)T = P(\epsilon_i)$, where $\sum_i P(\epsilon_i) = 1$. The collection of the causal states together with the transition probabilities define an *ϵ-machine.*

It is proven [13] that the ϵ-machine is

- a *sufficient* statistic, that is it contains the complete statistical information about the data;
- a *minimal sufficient* statistic, therefore the causal states can not be subdivided into smaller states;
- a *unique minimal sufficient* statistic, any other one simply re-labels the same states.

A.2. Symbolisation

Without any loss of dynamical information, an n-dimensional continuous trajectory of a dynamical system can be converted to an $(n-1)$-dimensional map using the Poincare section. At the locations where the trajectory pierces the Poincare section surface the points of the map are generated, thus sampling the continuous signal at discrete time moments. However, the dynamics of the map is equivalent to the original signal only if the full-dimensional phase space trajectory is considered. For molecular signals when the 3-dimensional configuration (or velocity) trajectory of one atom (or higher dimensional for a group of atoms) is analysed the Poincare map is undefined. However, a similar approach can be used to naturally sample the roughly periodic signal of molecular systems.

To discretise the three-dimensional velocity trajectories of individual atoms of the molecular system we used its intersections with the xy plane. For hydrogen water atoms, for example, the average time interval between the intersections was equal to 0.032 ps. Very conveniently it roughly corresponds to the first minimum on the autocorrelation function, obeying the general rule for time sampling of signals. The resulting two-dimensional points approximately uniformly cover the area and form a centrally-symmetric distribution of points, Fig. 5.

In order to convert the trajectory map into a sequence of symbols from a finite alphabet, an appropriate partitioning of the continuous space is required. A natural choice for such partitioning is the generating partition (GP) [14] that has the property of a one-to-one correspondence between the continuous trajectory and the generated symbolic sequence. That is, all information is retained after the symbolisation.

Consider a dynamical system $\mathbf{x}_{i+1} = \mathbf{f}(\mathbf{x}_i), \mathbf{f} : M \to M$ and a finite collection of disjoint open sets $\{B_k\}_{k=1}^K$, partition elements, such that for their closures $M = \cup_{k=1}^K \bar{B}_k$. Given an initial condition $\mathbf{x}_0$, the trajectory $\{\mathbf{x}_i\}_{i=-n}^n$ defines a sequence of visited partition elements $\{B_{\mathbf{x}_i}\}_{i=-n}^n$ or $\{s_i\}_{i=-n}^n$, where s_i are symbols from the alphabet that mark the elements where $\mathbf{x}_i \in B_i$. For a generating partition the intersection of all images and pre-images of these elements is, in the limit $n \to \infty$, a single point: $\cap_{i=-n}^n \mathbf{f}^{(-i)}(B_{\mathbf{x}_i})$.

This elegant mathematical construct has two disadvantages when applied to realistic molecular signals. First, an algorithm for calculating a GP in a general case is unknown. Second, it is shown for simple tent maps [15] that the values of statistical complexity for different GPs of the same system are different (a system can have many GPs, not to confuse with the uniqueness of a symbolic representation of a trajectory for a given GP).

Recently methods for finding approximations for GP are reported. The method from [16] is shown to reproduce GP for known systems and can treat multi-dimensional observed time-series data. The results of the application of this method to our velocity data using 2, 3, 4, and

5 partitions are shown in Fig. 5. For all cases the resulting approximations to GP are centrally symmetric (probably, because of the central symmetry of the data points distribution). Thus, for our signals we used centrally symmetric partitions in all subsequent calculations.

Summarising, in converting the three-dimensional molecular trajectories into symbolic sequences we, first, built a two-dimensional map by finding the intersections of the trajectory with the xy-plane and, second, assigned a symbol to each point of the map depending to what segment of the partition the point belongs (Fig. 5).

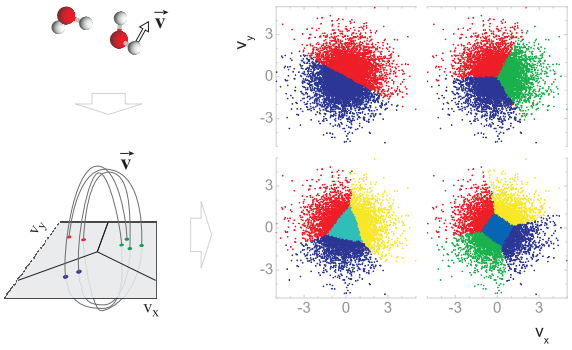

Figure 5. The process of converting the continuous atomic velocity signal **v** into symbolic sequence. On the right the symbolisation with 2, 3, 4, and 5 symbols are shown

A.3. Computational Mechanics produces consistent results

Two parameters of the algorithm should be set in calculating C_μ of a signal of given length, the alphabet size K and the length l of the histories s^- used by the ϵ-machine reconstruction algorithm `CSSR`.

The dependence of C_μ on both parameters is shown in Table 1. The convergence with l is excellent, so that for $l \geq 6$ the algorithm produces almost identical results. Reliable results for large alphabet sizes K are more difficult to obtain because for higher K much longer signals are required. This explains the somewhat increased values of C_μ for $K = 5$ in Table 1.

Table 1
Statistical Complexity C_μ vs. the length of histories l ($K = 3$) and the alphabet size K ($l = 9$) for bulk water hydrogen velocity 60 ns long signal

l	C_μ	K	C_μ
2	3.17	2	5.24
3	4.75	3	7.90
4	6.11	4	8.21
5	7.31	5	8.65
6	7.95		
7	8.15		
8	8.21		
9	8.29		
10	8.37		

Varying the position of the Poincare section plane along the z axes did not lead to any change in the results. The effect of various partitionings of the continuous space has been checked by applying non-symmetric (same as symmetric but shifted along the x and y axes) partitions. In all cases this resulted in lower values of C_μ. Any variants of centrally symmetric partitioning produced identical results.

REFERENCES

1. B. M. Ladanyi, M. S. Skaf, Computer simulation of hydrogen-bonding liquids, Annu. Rev. Phys. Chem. 44 (1993) 335–368.
2. A. M. Berezhkovskii, G. Sutmann, Time and length scales for diffusion in liquids, Phys. Rev. E 65 (2002) 060201.
3. R. Balescu, Statistical Dynamics, Matter Out of Equilibrium, Imperial College Press, 2000.
4. J. P. Crutchfield, K. Young, Inferring statistical complexity, Phys. Rev. Lett. 63 (2) (1989) 105–108.
5. J. P. Crutchfield, K. Young, Computation at the onset of chaos, in: by W. Zurek (Ed.), Entropy, Complexity, and Physics of Information, SFI Studies in the Sciences of Complexity, VIII, Addison-Wesley, Reading, Massachusetts, 1990.

6. J. P. Crutchfield, The calculi of emergence: computation, dynamics and induction, Physica D 75 (1-3) (1994) 11–54.
7. H. J. C. Berendsen, J. Postma, W. van Gunsteren, J. Hermans, Interaction model for water in relation to protein hydration, in: B. Pullman (Ed.), Intermolecular Forces, D. Reidel Publishing Company, Dordrecht, 1981, pp. 331–342.
8. D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. C. Berendsen, Gromacs: Fast, flexible and free, J. Comp. Chem. 26 (2005) 17011718.
9. H. J. C. Berendsen, Transport properties computed by linear response through weak coupling to a bath, in: M. Meyer, V. Pontikis (Eds.), Computer Simulations in Material Science, Kluwer, 1991, p. 139155.
10. C. R. Shalizi, K. L. Shalizi, Blind construction of optimal nonlinear recursive predictors for discrete sequences, in: M. Chickering, J. Halpern (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference, AUAI Press, Arlington, Virginia, 2004, pp. 504–511.
URL http://arxiv.org/abs/cs.LG/0406011
11. D. Nerukh, V. Ryabov, R. C. Glen, Complex temporal patterns in molecular dynamics: a direct measure of the phase space exploration by the trajectory at macroscopic time scales, Physical Review E.
12. F. Ritort, Single-molecule experiments in biological physics: methods and applications, Journal of Physics: Condensed Matter 18 (32) (2006) R531–R583.
URL http://stacks.iop.org/0953-8984/18/R531
13. C. Shalizi, K. Shalizi, R. Haslinger, Quantifying self-organization with optimal predictors, Physical Review Letters 93 (11) (2004) 118701–1 –118701–4.
14. S. Wiggins, Introduction to applied nonlinear dynamical systems and chaos, Springer, New York, 1990.
15. O. Gornerup, K. Lindgren, personal communication (2006).
16. M. Buhl, M. B. Kennel, Statistically relaxing to generating partitions for observed time-series data, Physical Review E 71 (2005) 046213.