On the question of calculating the free energies of biomolecular systems: how much of phase space is actually explored?

Dmitry Nerukh*
Unilever Centre for Molecular Informatics, Department of Chemistry,
Cambridge University, Cambridge CB2 1EW, UK

(Dated: April 3, 2009)

A novel statistical analysis of Molecular Dynamics generated trajectories is applied to various bulk liquids and a peptide in water. The analysis provides unique information on the *full dimensional* trajectory. In particular, it demonstrates that the phase space exploration is a *very slow process* that has the time scale of hundreds of nanoseconds even in bulk water and argon. Most importantly, the areas of the phase space visited at these times are *different*, in contrast to the commonly assumed uniform random search process. For a 21-residue peptide in explicit water it has been found that the peptide exhibits nanoseconds long periods that significantly differ in the rates of the phase space exploration. During these periods the rates remain the same but different from other periods and from the phase space covering rate in water.

I. INTRODUCTION

Many methods are developed for correct calculation of the free energy of molecular systems from Molecular Dynamics (MD) simulation (for a few recent examples see [1–5]). The most difficult problem here is a correct sampling of the phase space of the molecular system. Since the molecular dynamical system is extremely high dimensional, its phase space can not be sampled exhaustively in any feasible MD simulation. Thus, various non-trivial methods have to be used that provide a statistically correct coverage of the phase space areas involved in the chemical process under investigation.

Sampling of conformational space (that is coordinates only, no information on the momenta is used) is the subject of extensive investigation in the field of bio-molecular modelling. A good review can be found at [6]. An in-depth study [7] presents non-trivial results that demonstrate that (i) the sampling depends crucially on the quality of the forcefield of the simulation and (ii) the simulated physical chemistry characteristics of studied systems significantly depend on the quality of sampling. This includes to a large extend the reaction rates that, in turn, are defined by the "rare events" in the dynamics of the system [8].

Recently the question of the phase space sampling has attracted researchers' attention, at least it has been recognised that the issue is important both technically (in bio-molecular simulations) and conceptually. Perhaps, the most widely investigated area, besides the free energy calculation, is various methods of *artificial increasing* the phase space area explored by the system in the simulation.

These include many modifications of molecular dynamics aimed at accelerating the folding of proteins. In the area of bio-molecular simulations one of the most widely used method is Replica-Exchange Molecular Dynamics [9–12]. In this method several Molecular Dynamics simulations of the same system are run at different temperatures in parallel and the simulations periodically exchange temperatures. The range of temperatures allow the system to explore more of the phase space than is normally done in standard Molecular Dynamics. Another class of method for encouraging the conformational changes of the bio-molecule is Accelerated Molecular Dynamics or Hyperdynamics methods [13–15]. Here the change of the phase space exploration is achieved through the modification of the potential energy of the system. An extra term is added at the values of the energy below a given threshold. This reduces the energy barriers between the states and results in faster phase space exploration. A method designed specifically for peptides and proteins systems uses the construction of a Markov model of conformational transitions [16, 17]. The approach divides the simulation into independent parts that can be run on independent computers. However, unlike Replica-Exchange method the simulations are run at the same temperature. This technique has been pioneered in particular in the Folding@Home project and reached unprecedented scale of computation involving hundreds of thousands of computers.

A serious problem with these techniques is that they change the system (the forcefield or the dynamics) uncontrollably, thus risking to alter the native state of the protein. Indeed, as we have recently shown [18, 19], such changes can lead to not only a wrong folding state but also to a meaningless folding time. This is precisely because the system is artificially forced to explore the phase space areas that are different from the areas visited by the original, non-modified system. Therefore, the development of acceleration schemes for the folding MD simulations *that explore the phase space correctly* is of particular importance.

Despite all these works, there are very few investigations that focus on the sampling of the phase space directly. These are mostly the ones that study the reaction rates and associated events of the potential barrier cross-

*Electronic address: dn232@cam.ac.uk

ing (and recrossing). The authors of [20] concentrate on long pathways and find that the sampling provided by the "natural" trajectory (that is without an artificial acceleration) is not enough for obtaining robust statistics on the reacting trajectories. However, only the natural sampling provides correct data for calculating the reaction rates. As an example of recent new methodologies of robust techniques of correct phase space sampling can be the publication [21].

A crucially important assumption is utilised in all these methods: the MD trajectory is believed to sample the phase space *randomly*, at least at the time scale of several picoseconds. In other words, if we consider the points along the trajectory with the time step of the order of several picoseconds and longer, they are assumed to be completely equivalent to a purely random process and the probability of different points are exclusively defined by the energy differences at the points. This view seems to be supported by the standard autocorrelation analysis that shows that at these times there is no correlations in molecular signals, thus the signal is equivalent to a random process.

We would like to emphasize here the difference between "random" and "chaotic". The former is indistinguishable from the latter in terms of two point time correlations. Moreover, the molecular systems are commonly assumed to be highly chaotic and for this reason, at the time scales of several picoseconds, statistically equivalent. We provide more details on this in our previous publications [22, 23].

Another important consideration here is a specific nature of the free energy calculations. In sharp contrast to other dynamical characteristics, such as, for example, dynamical correlation functions or structure factors, it is important to consider *multi-point time correlations* here. The majority of the common characteristics are defined using only *two-point time correlations.* These could indeed show good statistics on a relatively short time scales. However, the inclusion of the critically important for the free energy calculations multi-point statistics makes the situation fundamentally different. Naturally, the latter is not an abstract academic exercise, but a practically crucial calculation since the free energy defines experimentally measurable reaction rates.

This approach leads to the "folding funnel" concept when applied to the process of protein folding. The concept introduces "non-randomness" at the level of transitions between the local minima on the funnel. The dynamics within the minima is still regarded purely random. Therefore, all useful information is contained in the free energy surface and indeed, the major efforts in this area are devoted to the analysis of the energy surfaces.

There are two immediate problems with this approach. First, it is not possible to investigate the whole free energy surface, it is infeasibly large. Second, the real molecular trajectory does not have enough time to sample the whole surface, nevertheless it finds the correct sequence of the minima that leads to the native conformation. Thus, elucidating the origin of this non-random phase space covering *directly* from the dynamic trajectory (not from the analysis of the potential energy of the system) is an important problem.

More specifically, the questions that we try to answer in this work are (i) how much of the phase space is actually covered by the molecular trajectory if it is allowed to evolve under the "true" dynamics (without any artificially introduced randomness) according to the equations of motion; (ii) at what times can we consider that enough phase space has been sampled by the trajectory? To answer these questions we use a sophisticated statistical analysis of the MD trajectories.

We show that our analysis provides information about the evolution of the full dimensional phase space trajectory of the system. It is possible to obtain very detailed information about the whole dimensional trajectory by analysing low dimensional (macroscopic) observables of the system such as, for example, individual atom's velocity, coordinate, or system's instant temperature. We demonstrate that the trajectory explores the phase space very slowly, at the time scale of hundreds of nanoseconds even for such homogeneous molecular systems as bulk water and argon. When applied to a peptide system our approach reveals long periods (dozens of nanoseconds) when the molecule is at "dynamical frustration", that is it does not explore other areas of the phase space.

Thus, we show that the commonly accepted random character of the phase space exploration generally does not hold. It is to a large extent defined by the complex dynamics of the molecular system, not just the Boltzmann distribution of the phase spaced density. Therefore, great care must be taken when applying methods that modify the way the phase space is explored by the molecular system.

II. THE METHOD

Molecular trajectory obtained in the simulation experiment is a series of $2N$ dimensional phase space points $\mathbf{q}_i$, where N is the number of degrees of freedom of the system, i.e. the number of atoms minus various constrains such as fixed bond lengths, angles, etc. N is of the order of several thousands for realistic MD simulations. Thus, the molecular trajectory is a very high dimensional object. The points are generated by the system along the trajectory at fixed time moments (Fig. 1).

To analyse the data almost always low dimensional observables (macro-observables) are considered, for example a velocity of an atom $\mathbf{v}$. This is a projection of the full dimensional trajectory onto a low dimensional observable, in this case $\mathbf{v}$. An important question is: by analysing the macro-observable what can we say about the whole dimensional phase space trajectory? More specifically, can we obtained any information on how much of the whole dimensional phase space is covered by a piece of trajectory?

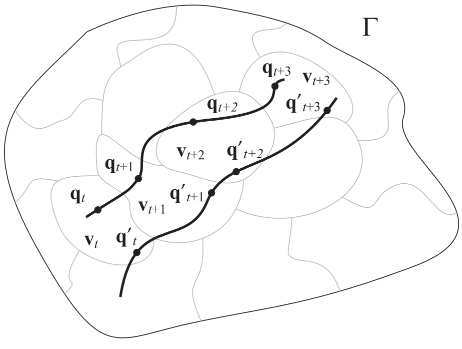

FIG. 1: Illustration of the degeneracy of the macro-observable projection of the full-dimensional phase space trajectory. The same sequence of the observable (the velocity) $\{\mathbf{v}_t\mathbf{v}_{t+1}\mathbf{v}_{t+2}\mathbf{v}_{t+3}\}$ is generated by two different pieces of the phase space trajectory $\{\mathbf{q}_t\mathbf{q}_{t+1}\mathbf{q}_{t+2}\mathbf{q}_{t+3}\}$ and $\{\mathbf{q}'_t\mathbf{q}'_{t+1}\mathbf{q}'_{t+2}\mathbf{q}'_{t+3}\}$

A key point to realise in the context of this work is that this low dimensional projection of the phase space trajectory is degenerate, that is very many different realisations of the trajectory produce the same series of values of the low dimensional projection $\mathbf{v}$ (Fig. 1). This is caused by (i) the discrete time sampling of the trajectory, (ii) the finite tolerance of the measurements of $\mathbf{v}$, and (iii) the independence of $\mathbf{v}$ at each individual time moments from some other degrees of freedom, for example the positions and velocities of distant at those moments atoms. Therefore, the whole phase space Γ is partitioned into the areas such that on each of them the macroscopic observable $\mathbf{v}$ takes a unique value while the full dimensional points $\mathbf{q}_i$ can have different values (Fig. 1).

The values of the observable variables that we analyse are discrete and finite. In other words, we deal with a set of countable number of *symbols*. In the case of the computer floating point representation, for example, the number of symbols is large but limited and defined by the precision used in the simulation (single, double, etc). The finite precision of $\mathbf{v}$ results in a finite (but large) set of its possible values. However, it is easy to check that even a very coarse representation of $\mathbf{v}$ produces almost the same characteristics of the analysed molecular signals. Fig. 2 shows one of such characteristics, the common velocity autocorrelation function for a signal where the velocity coordinates are replaced by only three values in $\mathbf{v}_x$,$\mathbf{v}_y$, and $\mathbf{v}_z$, such that $\{x \equiv -1, \text{if } x < -1; x \equiv 0, \text{if } -1 \leq x < 1; x \equiv 1, \text{if } x \geq 1\}$, where x represent $\mathbf{v}_x$,$\mathbf{v}_y$, and $\mathbf{v}_z$. The total number of possible values of the resulting coarse grained vector is $3^3 = 27$, that is the signal can be represented by only 27 symbols. Nevertheless, the autocorrelation function of this signal is very similar to the original one, calculated from the double precision values of $\mathbf{v}$.

This representation of the dynamics in terms of symbols from a finite size alphabet is called "symbolic dynamics" and is the subject of the mathematical field with the same name [24]. We here show that this is a very useful framework that allows to make unexpected conclusions about the phase space trajectory of the molecular system.

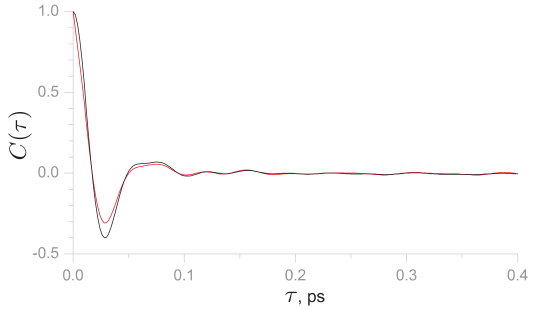

FIG. 2: Autocorrelation functions $C(\tau) \equiv \frac{1}{T}\sum_t^T \mathbf{v}_t \cdot \mathbf{v}_{t+\tau}$ for the original velocity of the hydrogen of bulk water (black) and the signal made of 27 symbols (red, see text or details)

A. The dynamics makes the partition finer

The evolution of the phase space points $\mathbf{q}$, sampled at times t, is governed by an operator $\mathbf{T}$: $\mathbf{q}_{t+1} = \mathbf{T}\mathbf{q}_t$. Because of the determinism the dynamics $\{\mathbf{q}_t\}$ forms a Markov chain. Considering an ensemble of such dynamical systems, denote a random variable representing the current microstate as $\mathbf{Q}$, that is a set of all possible values of the phase space points having probabilities generated by the dynamics $\mathbf{T}$.

A macroscopic observed variable A is a function f of the microstate $\mathbf{Q}$ (for example, the instantaneous temperature $\frac{1}{Nk}\sum_i m_i\mathbf{v}_i^2$, where N is the number of degrees of freedom, k is the Boltzmann constant, m_i are the atoms' masses, and $\mathbf{v}_i$ are their velocities). As discussed before, the function f partitions the phase-space Γ into mutually exclusive and jointly exhaustive sets, on each of which f takes a unique value. Denote the partition of Γ induced by f as $\mathcal{F}$. The observed process is $A_t = f(\mathbf{Q})$ and it is not necessarily Markovian (Fig. 3).

Now, what happens to this partition when the sequences of A_t are considered instead of the inidividual values of A? Take an observation at time t, A_t. The corresponding set of point in Γ is $\mathcal{F}_t$. For a sequence of two observations at the current and previous time moments the set of points is

$$\mathcal{F}_t \cap \mathbf{T}\mathcal{F}_{t-1}, \quad (1)$$

which is a refinement of the partition $\mathcal{F}$. This procedure can be repeated any countable number of times thus providing the refined partitions for the histories of the macro-observable A. Thus, the dynamics makes the initial partition induced by the macro-observable finer, the longer the sequence $\{A_t\}$ (the "history") the finer the partition generated by the sequence is.

B. Computational Mechanics [25] coarsens the partition

The next step is to apply a special statistic, called Computational Mechanics (CM) [25], to the observable A. The rigorous definition of CM is given in Appendix A. Here we provide the part of the approach necessary for answering the main question formulated in the Introduction.

All past A_i^- and future A_i^+ halves of bi-infinite sequences of the macro observable centred at times i are considered. Two pasts A_1^- and A_2^- are defined equivalent if the conditional distributions over their futures $P(A^+|A_1^-)$ and $P(A^+|A_2^-)$ are equal. A *causal state* $\epsilon(A_i^-)$ is a set of all pasts equivalent to A_i^-: $\epsilon_i \equiv \epsilon(A_i^-) = \{\lambda : P(A^+|\lambda) = P(A^+|A_i^-)\}$. 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*. The *Statistical Complexity* is the informational measure of the size of the ϵ-machine: $C_\mu = H[P(\epsilon_i)]$, where P are the probabilities of the causal states and H is the Shannon entropy of the distribution of a random variable ν, $H[P(\nu)] \equiv -\sum_\nu P(\nu)\log_2 P(\nu)$.

Thus, the essence of CM is in grouping the histories $\{A_t\}$ into causal states. In terms of the partitions of the phase space this corresponds to joining together the cells of Γ induced by the dynamics. Importantly, the new cells represent a Markovian process constructed from the observed process A_t by building the ϵ-machine on A. Now by the ϵ-machine definition the sequence of the causal states $\{\epsilon_t\}$ makes a Markov chain (Fig. 3).

C. The partition generated by Computational Mechanics is the most informative one

Shalizi and Moore [26] show that in this setting the Statistical Complexity has a clear physical meaning: it quantifies the amount of information contained in the new constructed macro-observable process $\{\epsilon_i\}$ about the microstate:

$$C_\mu = I[\mathbf{Q}; \epsilon], \tag{2}$$

where I is the mutual information between random variables X and Y: $I[X;Y] = H[X] - H[X|Y]$; and $H[X|Y]$ is a conditional entropy of X given Y: $H[X|Y] = -\sum P(X)\sum P(X|Y)\log_2 P(X|Y)$.

This is because the knowledge of the microstate would specify the macro observable precisely: $H[\epsilon|\mathbf{Q}] = 0$, because all histories contained in ϵ_t and the corresponding partition of $\mathbf{Q}$ would uniquely define the next state ϵ_{t+1} (the ϵ-machine definition). Using this and the equality $H[X] + H[Y|X] = H[Y] + H[X|Y]$ the equation (2) follows:

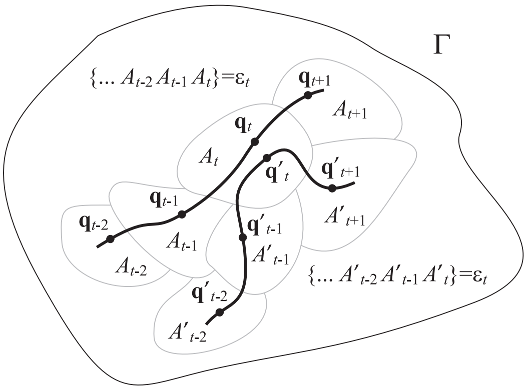
Γ
{... $A_{t\text{-}2}\,A_{t\text{-}1}\,A_t$}=$\varepsilon_t$
$\mathbf{q}_{t+1}$
A_{t+1}
$\mathbf{q}_t$
A_t
$\mathbf{q}'_t$
$\mathbf{q}'_{t+1}$
$\mathbf{q}_{t\text{-}1}$
$\mathbf{q}_{t\text{-}2}$
$A_{t\text{-}1}$
$\mathbf{q}'_{t\text{-}1}$
A'_{t+1}
$A_{t\text{-}2}$
$A'_{t\text{-}1}$
{... $A'_{t\text{-}2}\,A'_{t\text{-}1}\,A'_t$}=$\varepsilon_t$
$\mathbf{q}'_{t\text{-}2}$
$A'_{t\text{-}2}$

FIG. 3: Schematic illustration of the sequences used to define formula (2). Phase space points $\{\mathbf{q}\}$ and $\{\mathbf{q}'\}$ of two pieces of the trajectory form Markov sequences. The corresponding observation sequences A and A' are not Markovian since the same value A_t leads to different A_{t+1} and A'_{t+1} depending on the previous values A_{t-1} and A'_{t-1}. However, if both histories $\{\ldots A_{t-2}A_{t-1}A_t\}$ and $\{\ldots A'_{t-2}A'_{t-1}A_t\}$ belong to the same causal state ϵ_t than the next causal state ϵ_{t+1} is defined without knowing ϵ_{t-1}, thus making $\{\epsilon\}$ a Markov sequence.

$$\begin{aligned}
H[\mathbf{Q}|\epsilon] + H[\epsilon] &= H[\epsilon|\mathbf{Q}] + H[\mathbf{Q}] \\
H[\mathbf{Q}|\epsilon] + C_\mu &= H[\mathbf{Q}] \\
C_\mu &= H[\mathbf{Q}] - H[\mathbf{Q}|\epsilon] \\
C_\mu &= I[\mathbf{Q}; \epsilon].
\end{aligned}$$

Because of the properties of the ϵ-machine this is the maximal information that is possible to extract from the chosen macro-observable and the specified initial partition of it.

D. Three stages of symbolisation

Summarising, we have arrived at the phase space partition, obtained using three stages.

1. The observed macro-variable induces an initial (usually very coarse grained) partition of the phase space.
2. The cells of this partition are refined by the dynamics (1).
3. The refined cells are grouped by the process of ϵ-machine reconstruction, thus providing the final partition that is the minimal, unique, and most informative one.

E. Phase space exploration

The above reasoning is based on the assumption that the trajectory is infinitely long and covers the whole available phase space. Therefore, some estimations have to be done as for how long should the trajectory be in the simulation that would correctly represent the covering of the phase space. The first requirement is, obviously, that the number of histories, $\{A_t\}$, be enough that the cells (1) are populated with at least several histories each. Secondly, if new pieces of the trajectory passing through the same cells produce the same ϵ-machine that would signify that we have a long enough simulation.

When the trajectory gradually fills the phase space over longer times, there are two situations that can happen. Provided that we have a long enough trajectory that satisfies the requirements above, with time the ϵ-machine can change. This would mean that (i) the full dimensional trajectory explores previously not visited areas of the phase space and (ii) the information about the trajectory contained in the observable is different at these new areas of the phase space.

III. MOLECULAR SYSTEMS AND SIGNAL PROCESSING

Clusters of 3, 7, 15, and 52 water molecules in vacuum, bulk water (periodic boundary conditions) consisting of 392 or 878 SPC or SPC-E [27] molecules, and bulk argon (Lennard-Jones particles) were simulated using the `GROMACS` molecular dynamics [28] package. The temperature of the systems was kept constant at 300K using Berendsen [29] or Nose-Hoover [30] thermostats whose combination with various coupling constants was investigated. A sufficient equilibration was performed before collecting data for analysis.

We have chosen a 21-residue peptide $A_5(A_3RA)_3A$ from the review [31] where it is reported to fold in 0.8 μs on average. The forcefield for the simulations was GROMOS96 [32]. The peptide was solvated by 1658 SPC water molecules [27] and after proper minimisation of the system's energy was simulated for 0.5 μs. We have not reached the folded state, however, prolonged periods of the existence of β-sheet and α-helix motifs were recorded. The velocities of one of the water hydrogens, and of the nitrogens of the residues 1 and 3 were taken for the analysis (see Appendix B for the signal processing details).

As discussed, various molecular signals can be used for the analysis, any such signal is a function of the microstate $\mathbf{Q}$. For the studies reported in this paper we have chosen the velocities and coordinates of various atoms and the instantaneous temperature. As the initial partitioning (section II) we have used an approximation of the generating partition (see Appendix B) that divided the velocity space into three centrally symmetric sectors.

We have found (see Appendix C) that it is possible to obtain consistent results provided that the correlations in the molecular signal vanish at sub-picosecond times. The details of the procedure of the initial symbolisation are provided in Appendix B.

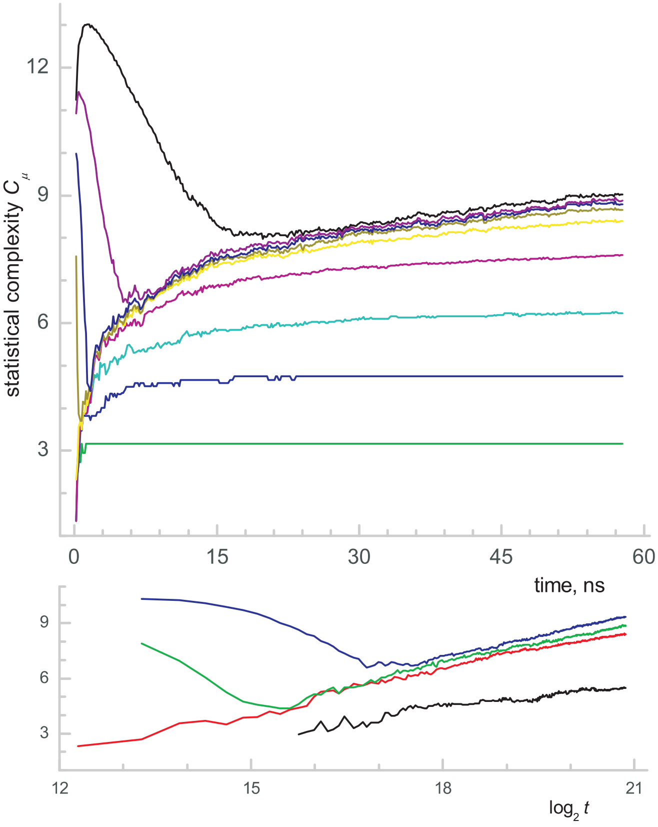

FIG. 4: Statistical complexity (top) and the logarithmic dependence of C_μ (bottom) on the number of data points N for the hydrogen velocity signal of bulk water at 300K. Top: the curves, from green to black, correspond to the values of the history length l from 2 to 10. Bottom: black, red, green, and blue lines correspond to the alphabet size $K = 2, 3, 4, 5$ respectively.

IV. RESULTS

There are three parameters that can be adjusted in the ϵ-machine reconstruction algorithm, `CSSR`: the size of the alphabet for symbolisation K, the length l of the histories $\{A_t\}$ used for the reconstruction, and the total length of the signal T. While C_μ practically converges for $l > 7$ and $K > 3$ (see Appendix C), it shows non-trivial dependence on the length of the signal (simulation time) at surprisingly long times, Fig. 4.

C_μ quickly increases and then decreases eventually settling on the $\log_2 T$-like curve at the times of ≈8 ns (this depends on l). The curves keep growing in exactly the same manner until the simulation times of 1 μs. 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 (see Appendix B 3 for the discussion on the requirements for the length of the signal). The number of causal states, n_{st}, at these values of T is very high and each state consists of only a few histories, that is the first requirement from section II E is not satisfied. This part of the curve is of little interest for

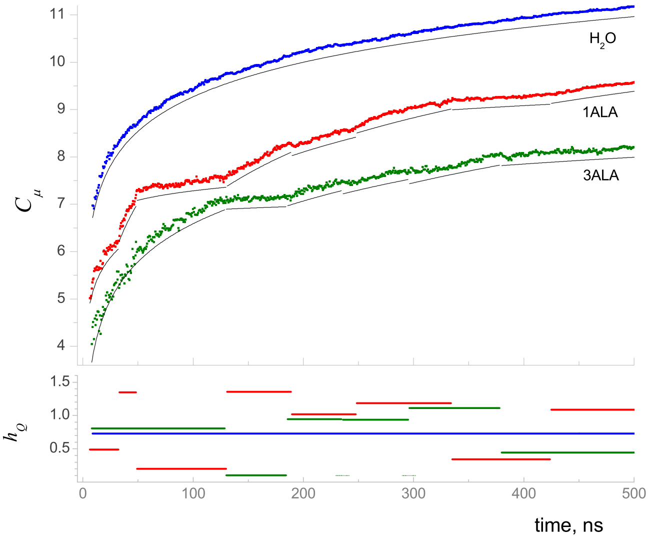

FIG. 5: Top: the trajectory length dependence of C_μ for: blue - water hydrogen, red - residue 1 nitrogen, green - residue 3 nitrogen. Bottom: the values of h_Q at the intervals of the constant growth

the present analysis and only the logarithmic part of the curves are discussed here and this represents the main phenomenon reported.

It should be emphasised that the length of the signal T here corresponds to extremely long times. In a molecular system such as bulk water no transient effects can be expected at these times, which is confirmed by the constant value of the Shannon entropy of the symbolic sequence itself ($H[P(A_i)] = 1.58490 \pm 0.00005$ for $t > 2$ ns), as well as unchanged velocity autocorrelation function.

While for the bulk water and other bulk liquids a perfect line on the $log_2 T$ - C_μ plot is observed, the peptide exhibits well pronounced periods with significantly different rates of the growth. Within one period the growth can still be satisfactory fitted with a line. Importantly, the changes between the periods are quite sharp such that the whole curve is divided into well separated parts, Fig. 5.

V. DISCUSSION

The analysis of the changes of C_μ with the length of the history l shows that starting from the length 6 the C_μ almost does not change, Fig. 4. This means that this length of the history makes the partition as fine as possible (see section II A). Further splitting of the phase space cells does not provide any more information - the following step of joining them by the causal states reverses the situation to the cell size approximately equal to that of $l = 6$.

The fact of the growing number of causal states with T is extremely interesting since it proves that the new phase space areas visited by the trajectory are *different from those visited before*. That is, even after several dozens of nanoseconds the trajectory keeps exploring new areas. Note that the phase space cells after stage 2 are exactly the same for all times, what changes with T is the process of coarsening, stage 3. The more data is available from the trajectory exploration, more information about the phase space is introduced and this information is different in different areas of the phase space.

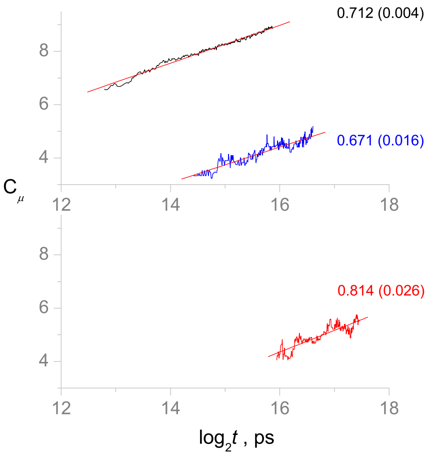

FIG. 6: h_Q values (indicated on the right) for various observables of bulk water: black - the hydrogen velocity, red - the oxygen velocity, and blue - the instantaneous temperature

The slope, h_Q, of the logarithmic part of $C_\mu(T)$ can serve as a measure of the rate with which the phase space of the system is explored:

$$C_\mu = a + h_Q \log_2 T$$

The value of this characteristic is a fundamental property of the system which does not seem to depend on the details of the simulation model, the number of the molecules (for bulk systems), and even on the temperature. However, it does depend on the chemical nature of the system, that is the values of h_Q are different for different chemical systems.

Another remarkable fact is that the rate of change of C_μ is *the same for very different macro-observables*. The complexity values were calculated for the velocities of the oxygen and hydrogen atoms and the instantaneous temperature in bulk water. The results are presented in Fig. 6 and clearly show a good agreement in the h_Q values.

h_Q values for other molecular systems differ from the values shown in Fig. 6. For example, 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. It does not seem to depend on the details of the simulation model or even temperature. The results for bulk water presented in Table I are comparable even though TIP3P water model and bulk water at high temperature produce slightly higher

TABLE I: h_Q values for various MD models and temperatures and for a set of independently simulated systems (see text).

MD model	temperature, K	h_Q
TIP3P	300	0.78
SPCE	300	0.69
SPC	275	0.70
SPC	300	0.71
SPC	380	0.76
set	300	1.21

values of h_Q.

The values of C_μ together with the corresponding values of h_Q for the protein system are shown in Fig. 5. Long periods of very slow changes of h_Q are clearly visible. These periods signify that the trajectory is most probably visits the same phase space areas, at least the areas that do not introduce new information into the velocity projection.

VI. CONCLUSIONS AND OUTLOOK

The application of a sophisticated statistical analysis, Computational Mechanics, to molecular trajectories shows that it is possible to extract a detailed information about the whole dimensional trajectory by analysing low dimensional observables such as velocities of atoms, their coordinates, or the instantaneous temperature.

Most interestingly, we demonstrate that the trajectory explores the phase space very slowly, at the time scale of hundreds of nanoseconds. The new areas of the phase space seen by the trajectory are different from the previously visited areas and carry different statistical information. We have also found that the latter is very different for various simple liquids and fundamentally different in complex self-organising systems such as a peptide in water.

These results show that when calculating the free energy from MD simulations the non-randomness of the phase space exploration has to be taken into account. The phenomenon can be related to experiments, the main difficulty here is the requirement of the signal from a single molecule, not an ensemble average, normally measured in the experiment. Good perspective in this connection is in recently very active areas of single molecular methodologies. The work on connecting our finding to these experimental techniques is our current activity.

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.

APPENDIX A: 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 [33] 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.

APPENDIX B: SIGNAL PROCESSING

1. Discretisation

Without any loss of dynamical information, an n-dimensional 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. 7.

2. Symbolisation

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) [34] 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 [35] 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 [36] 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. 7. 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.

3. Signal properties

In order to obtain statistically correct results the symbolic signal should be long enough to satisfy the following criteria [37].

It is demonstrated [38] that to consistently estimate the probabilities of symbolic subsequence of length l in a signal with entropy rate h (for blocks of symbols $s^l \equiv s_1, \ldots, s_l$ the entropy rate [39] of the entire infinite sequence is defined as $h \equiv \lim_{l \to \infty} H[P(s^l)]/l$) the length of the signal has to be at least 2^{hl}.

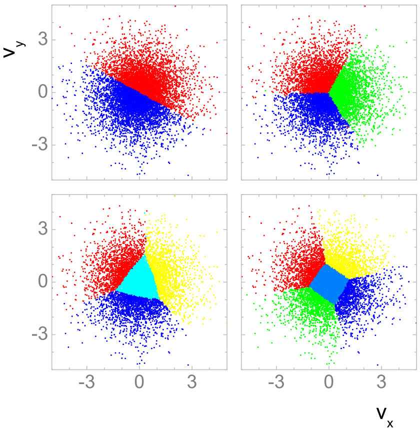

FIG. 7: Approximations for generating partitions obtained using the method by Buhl and Kennel [36] for the discretised hydrogen velocity for 2, 3, 4, and 5 partitions.

It is reasonable to require the length of the histories l to be such that the time interval covered by l symbols exceeds all correlations in the original signal. For the hydrogen velocities example from Fig. 4 $l = 0.2/0.032 \approx 6$. For the same data the entropy rate is $h = 1.56$ (for $K = 3$). Thus, the total sequence length should be longer than 657 symbols, which is definitely the case since our simulation times are of the order of nanoseconds, that is millions of data points.

From these considerations it also follows that only the signals with relatively short correlation times can be reliably analysed using computational mechanics. For example, if the correlation time exceeds ≈0.5 ps that would require more than 20 million points that corresponds to the times ≈700 ns and becomes problematic for atomistic MD simulations.

APPENDIX C: COMPUTATIONAL MECHANICS PRODUCES CONSISTENT RESULTS

Two parameters of the algorithm should be set in calculating C_μ of a signal of given length (we used a trajectory of 30 ns long, that is ≈1 million data points), 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 II. 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 the value of the entropy rate h is also high. Therefore, much longer signals are required. This explains the somewhat increased values of C_μ for $K = 5$ in Table II.

Varying the position of the Poincare section plane along the z axes did not lead to any change in the results.

TABLE II: Statistical Complexity C_μ vs. the length of histories l (total signal length is 30 ns, $K = 3$) and the alphabet size K (similar signal, $l = 9$) for pbc water hydrogen velocity 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		

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. This, we believe, serve as further evidence that the symmetric partition is a good approximation of GP.

[1] C. D. Christ and W. F. van Gunsteren, The Journal of Chemical Physics **126**, 184110 (pages 10) (2007), URL `http://link.aip.org/link/?JCP/126/184110/1`.

[2] C. D. Christ and W. F. van Gunsteren, Journal of Chemical Physics **128** (2008).

[3] D. Wu and D. A. Kofke, The Journal of Chemical Physics **123**, 054103 (pages 10) (2005), URL `http://link.aip.org/link/?JCP/123/054103/1`.

[4] D. Wu and D. A. Kofke, The Journal of Chemical Physics **123**, 084109 (pages 10) (2005), URL `http://link.aip.org/link/?JCP/123/084109/1`.

[5] L. Zheng, I. O. Carbone, A. Lugovskoy, B. A. Berg, and W. Yang, J Chem Phys **129**, 034105 (2008).

[6] M. Christen and W. F. Van Gunsteren, Journal of Computational Chemistry **29**, 157 (2008).

[7] E. J. Sorin and V. S. Pande, Biophysical Journal **88**, 2472 (2005).

[8] E. E. Borrero and F. A. Escobedo, J Chem Phys **129**, 024115 (2008).

[9] Y. Sugita and Y. Okamoto, Chemical Physics Letters **314**, 141 (1999).

[10] X. Periole and A. E. Mark, Journal of Chemical Physics **126**, 11 (2007).

[11] A. Baumketner and J. E. Shea, Theoretical Chemistry Accounts **116**, 262 (2006).

[12] K. P. Ravindranathan, E. Gallicchio, R. A. Friesner, A. E. McDermott, and R. M. Levy, Journal of the American Chemical Society **128**, 5786 (2006).

[13] D. Hamelberg, J. Mongan, and J. A. McCammon, Journal of Chemical Physics **120**, 11919 (2004).

[14] A. F. Voter, Physical Review Letters **78**, 3908 (1997).

[15] A. F. Voter, Journal of Chemical Physics **106**, 4665 (1997).

[16] N. Singhal, C. D. Snow, and V. S. Pande, Journal of Chemical Physics **121**, 415 (2004).

[17] G. Jayachandran, V. Vishal, and V. S. Pande, The Journal of Chemical Physics **124**, 164902 (pages 12) (2006), URL `http://link.aip.org/link/?JCP/124/164902/1`.

[18] C. H. Jensen, D. Nerukh, and R. C. Glen, The Journal of Chemical Physics **128**, 115107 (2008).

[19] C. H. Jensen, D. Nerukh, and R. C. Glen, The Journal of Chemical Physics **129**, 225102 (pages 6) (2008), URL `http://link.aip.org/link/?JCP/129/225102/1`.

[20] M. Grunwald, C. Dellago, and P. L. Geissler, The Journal of Chemical Physics **129**, 194101 (pages 8) (2008), URL `http://link.aip.org/link/?JCP/129/194101/1`.

[21] J. Kuipers and G. T. Barkema, J Chem Phys **128**, 174108 (2008).

[22] D. Nerukh, V. Ryabov, and R. C. Glen, Physical Review E **77**, 036225 (2008).

[23] D. Nerukh, Chemical Physics Letters **457**, 439 (2008).

[24] D. Lind and B. Marcus, *An introduction to symbolic dynamics and coding* (Cambridge University Press, New York, NY, USA, 1995), ISBN 0-521-55900-6.

[25] J. P. Crutchfield and K. Young, Phys. Rev. Lett. **63**, 105 (1989).

[26] C. R. Shalizi and C. Moore, Studies in History and Philosophy of Modern Physics **submitted** (2003), URL `http://arxiv.org/abs/cond-mat/0303625`.

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

[28] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, J. Comp. Chem. **26**, 17011718 (2005).

[29] H. J. C. Berendsen, in *Computer Simulations in Material Science*, edited by M. Meyer and V. Pontikis (Kluwer, 1991), p. 139155.

[30] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[31] J. Kubelka, J. Hofrichter, and W. A. Eaton, Current Opinion in Structural Biology **14**, 76 (2004).

[32] W. Scott, P. Hunenberger, I. Tironi, A. Mark, S. Billeter, J. Fennen, A. Torda, T. Huber, P. Kruger, and W. van Gunsteren, J. Phys. Chem. A **103**, 3596 (1999).

[33] C. Shalizi, K. Shalizi, and R. Haslinger, Physical Review Letters **93**, 118701 (2004).

[34] S. Wiggins, *Introduction to applied nonlinear dynamical systems and chaos* (Springer, New York, 1990).

[35] O. Gornerup and K. Lindgren, personal communication (2006).

[36] M. Buhl and M. B. Kennel, Physical Review E **71**, 046213 (2005).

[37] C. Shalizi, personal communication (2006).

[38] K. Marton and P. C. Shields, Annals of Probability **23**, 960 (1994).

[39] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Champaign-Urbana, 1962).