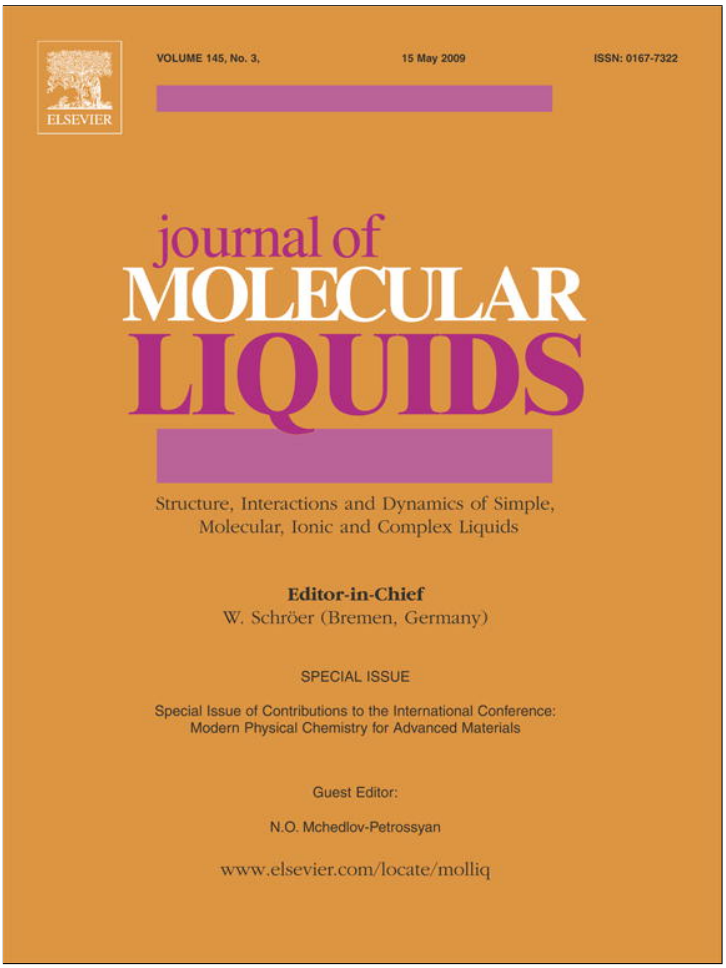
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Journal of Molecular Liquids 145 (2009) 139–144

Contents lists available at ScienceDirect

Journal of Molecular Liquids

journal homepage: www.elsevier.com/locate/molliq

Dynamical frustration of protein's environment at the nanoseconds time scale

Dmitry Nerukh *

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

ARTICLE INFO

Available online 25 June 2008

Keywords:
Dynamical frustration
Protein's environment
Phase space trajectory
Molecular dynamical system
Computational mechanics
Statistical complexity
Protein folding

ABSTRACT

A 21-residue peptide in explicit water has been simulated using classical molecular dynamics. The system's trajectory has been analysed with a novel approach that quantifies the process of how atom's environment trajectories are explored. The approach is based on the measure of Statistical Complexity that extracts complete dynamical information from the signal. The introduced characteristic quantifies the system's dynamics at the nanoseconds time scale. It has been found that the peptide exhibits nanoseconds long periods that significantly differ in the rates of the exploration of the dynamically allowed configurations of the environment. During these periods the rates remain the same but different from other periods and from the rate for water. Periods of dynamical frustration are detected when only limited routes in the space of possible trajectories of the surrounding atoms are realised.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Despite intensive research the protein folding problem remains largely unsolved. While the commonly accepted picture of "folding funnel" explains the overall behaviour of the system during folding, dynamically it is unclear what drives the molecular trajectory through the structural changes leading to the native state. Two features of molecular dynamical systems present major difficulties: its extremely high dimensionality and nanoseconds time scale between the major configurational changes of the protein molecule (the formation of the folding motifs such as α-helix and β-sheet). The local dynamics of the system (the only source of complicated folding behaviour) is commonly described using molecular parameters such as correlation times and transport coefficients with characteristic time scale of the order of few picoseconds. Thus, new methodologies that provide information about the behaviour of the high dimensional molecular trajectory at the nanoseconds time scale are desirable.

Classical molecular dynamics is a tool that is capable of simulating realistic protein systems in water at nanoseconds times. Currently computational power is enough to simulate small fast folding peptides at times up to complete folding. Therefore, it is now possible to obtain trajectories of the system with virtually any precision and details. This provides a unique opportunity to develop new conceptual methodologies for studying the dynamics of the system at the time scale of the conformational changes and up to folding in details inaccessible in experiment.

Molecular dynamics has long been used to analyse the dynamics of molecules by utilising, among other characteristics, various autocorrelation functions and diffusion constants. In relation to protein folding it has been found that water molecules around protein exhibit anomalous diffusion and behave like water at a lower temperature than the bulk water [1–3]. The velocity autocorrelation functions of protein atoms have been analysed and related to experimental spectroscopic data [4].

In all these studies, however, the time covered by the dynamical quantities does not exceed few tens of picoseconds. Moreover, to the best of our knowledge, all known dynamical characteristics of molecular systems that quantify local dynamics (not the gross quantities such as the parameters of the folding motifs, gyration ratio, etc. that are the *result* of the local dynamics) reach their limiting values at these times and remain unchanged at longer times. For example, deviations from normal diffusion have been detected by the analysis of higher moments of the mean square displacement in water [5,6] at times longer than the commonly accepted few picoseconds. However, the diffusion reaches its limiting value at times above ≈100 ps [5,7]. Nevertheless, despite this absence of correlations at longer times the local dynamics leads to the emergence of non-trivial structure at the time scale several orders of magnitude longer than the characteristic time of the common descriptors.

In the present study we introduce a fundamentally new dynamical characteristic, based on the information theoretical approach Computational Mechanics [8–10]. We show that the methodology provides quantitative information on the process of how the space of allowed trajectories of the neighbouring atoms is explored. The time spawned by our measure is of the order of nanoseconds, that is the characteristic time of elementary structural changes (formation and destruction of the basic folding motifs such as helices, sheets, turns, etc.) in realistic proteins during folding. The measure has been shown to reveal time correlations in molecular signals at the hundreds of picoseconds time scale and even longer [11]. It quantifies the rate with which the trajectory of the system explores the allowed areas in the phase space [12].

* Fax: +44 1223 763076.
E-mail address: dn232@cam.ac.uk.

0167-7322/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.molliq.2008.06.012

Author's personal copy

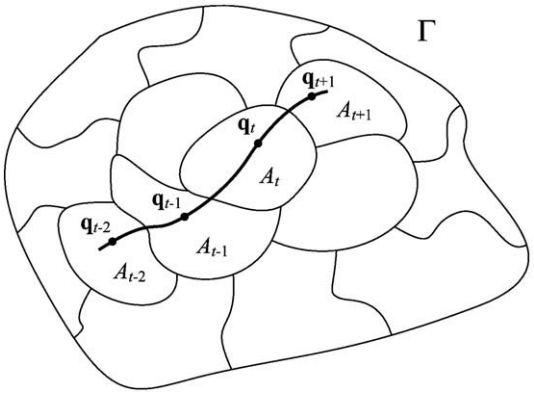


Fig. 1. Schematic illustration of the mapping of a full-dimensional phase space trajectory $\mathbf{q}_i$ to a macroscopic observable A_i. The phase space Γ is partitioned such that in each partition the value of A is the same.

We have simulated a 21-residue peptide in explicit water that is known to quickly form an α-helix. We have found that while water molecules' environment is explored with the same rate as in the bulk water [11] during the whole time of simulation, the protein atoms exhibit long periods that differ in the rate of the exploration. Because of the matching time scale we hope that the introduced measure can be related to the structural changes occurring in the peptide.

2. The idea

Consider a molecular trajectory in the full-dimensional phase space. With time it generates $2N$-dimensional points $\mathbf{q}_i \equiv (\mathbf{x}_i, \mathbf{p}_i)$ (Fig. 1), where N is the number of the system's degrees of freedom and $\mathbf{x}_i$, $\mathbf{p}_i$ are the coordinates and momenta of the atoms. N is of the order of several thousands even in MD simulations, and of the order of Avogadro number for real systems. In order to extract any sensible information from the series $\{\mathbf{q}_i\}$ a function has to be chosen that converts the full-dimensional points to a low-dimensional observable $A_i = f(\mathbf{q}_i)$. Because of the low-dimensionality of A and the finite precision of measurements, different $\mathbf{q}_i$ can be mapped to the same values of A. Therefore, the function f partitions the phase space Γ into mutually exclusive and jointly exhaustive sets, on each of which f takes a unique (up to the tolerance) value (Fig. 1).

For example, the velocity of one of the atoms $\mathbf{v}$ can be taken as such a function. The velocity is the function of all phase space variables in the sense that its value depends on the coordinates and momenta of all other atoms (taking into account the past values of the latter, see the details at the end of this section). It is easy to see that in this case different values of $\mathbf{q}$, when the differences are in the positions or velocities of distant atoms, correspond to the same values of $A \equiv \mathbf{v}$, that is the velocity of the chosen atom.

Thus, we would like to analyse a low-dimensional signal $\{A_i\}$. The analysis is normally done in the form of a statistic on the series, considering A as a random variable, that is a stochastic process (a deterministic signal is a limiting case of the stochastic process with the probabilities being the δ-functions). Elaborating the example with the velocities the observable can be chosen to be the scalar product of the velocity values at times t and $t+\tau$: $C_A(\tau) \equiv \mathbf{v}_t \cdot \mathbf{v}_{t+\tau}$.[1] If the statistic is calculated as a simple time average we obtain the velocity autocorrelation function $C(\tau) \equiv \frac{1}{T}\sum_t^T \mathbf{v}_t \cdot \mathbf{v}_{t+\tau}$, where T is the total lumber of points in the signal.

In Computational Mechanics [8–10] the statistic is built on the histories (pasts) of the observable $\{A_t^- \equiv \ldots A_{t-2}A_{t-1}A_t\}$ in such a way that they are grouped into so-called "causal states". The criteria of grouping is defined on the future sequences (futures) that follow each history. Two pasts A_i^- and A_j^- belong to one causal state if they produce the same (statistically) futures. In terms of probabilities two pasts are defined equivalent if the conditional distributions over their futures $P(A^+|A_i^-)$ and $P(A^+|A_j^-)$ are equal. In this way all possible histories of the signal are distributed between the causal states. Moving from point to point in the observed series $\{A_i\}$ is converted to the transitions from one causal state to another. There is a number of very useful fundamental properties of this representation of the signal that are described in Appendix A. Perhaps the most important property of the formalism in the scope of the present study is the completeness of the dynamical information extracted from the signal. This is in contrast to almost all dynamical characteristics used to analyse molecular dynamics (autocorrelation function, for example, is a very crude, two-point linear characteristic).

Consider two signals of the observable $\{A_i\}$, $i=0\ldots T_1$ and $i=0\ldots T_2$. Let's choose $T_{1,2}$ such that they are significantly longer than the characteristic time of the analysed dynamical quantity. For the velocity autocorrelation example the characteristic time can be the value of i when the correlations become essentially zero. For the Computational Mechanics case this can be the length of the histories A^- at which the causal states structure does not change (see Section 3).

In this setting the phase space trajectory passes many times through the same partitions A_i accumulating the statistics on the same sequence of the observable (Fig. 2). The statistics produced by the two signals would be different if the values of $T_{1,2}$ are not high enough, that is new passes of the trajectory alter the probabilities of the observable sequences. At some values $T_{1,2} > T'$ the statistics would become essentially the same. This limiting statistic would produce a limiting value of the analysed dynamical quantity (in the autocorrelation function example the function would converge to its "true" value). To the best of our knowledge for all common dynamical characteristics of molecular systems T' does not exceed few tens of picoseconds.

We have found that in the case of Computational Mechanics, the causal states structure *never converges* at least at the lengths of feasible MD simulations. From the phase space trajectory point of view this means that the system produces different futures for the same pasts at all observed in the simulation times (Fig. 2). 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. Extensive tests illustrating this are provided in [12] and summarized in Appendix B.

Similar behaviour can be observed for, for example, the mean square displacement of an atom $\langle x^2(t)\rangle$, where x is an atom's coordinate. This quantity also diverges with time. However, there is a fundamental difference in the case of Computational Mechanics: non-trivial behaviour at all times is observed, whereas the diffusion constant characterising the displacement quickly reaches its limiting value.

The differences in the dynamics of $\mathbf{v}$ are completely defined by the dynamics of the environment of the atom, that is the coordinates and momenta of the neighbouring atoms. Indeed, the time evolution of the velocity is defined by the Newton equation $\dot{\mathbf{v}} = -\frac{1}{m}\mathbf{F}$, where the force $\mathbf{F}$ derived from the interatomic interaction potential V is a function of the surrounding atoms' coordinates $\mathbf{x}_i$: $\mathbf{F} = -\nabla V \equiv f(\mathbf{x}_i)|_{i=1..N}$. When considering the *histories* of the velocity values $\{\mathbf{v}_t\}$ they become the functions of the *histories* of the neighbouring atoms' coordinates and momenta (Fig. 3).

The number of possible combinations of the trajectories of the atoms influencing the atom under consideration is extremely large. Therefore, a long time is required for exhaustively sampling all of them. That is the statistic on them, quantified by the ϵ-machine, would change as more and more dynamically allowed combinations of the trajectories are realised in the simulation. In this sense the changes in the ϵ-machine quantify how the space of neighbourhood atoms'

[1] More precisely, in this case the observable is more involved: it is not a number, but a τ-dimensional vector of the products $C_A \equiv \{\mathbf{v}_t \cdot \mathbf{v}_{t+j} | j=0\ldots\tau\}$ for each time t.

Author's personal copy

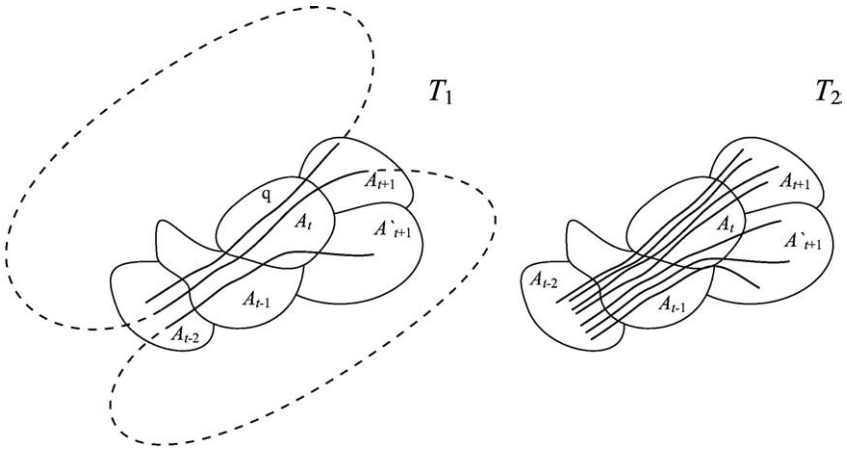


Fig. 2. Two signals $\{A_i\}$ of different lengths $T_{1,2}$ generated from the same molecular trajectory. The trajectory **q** passes through the same partitions of the macro-observable A thus accumulating the statistic on the history $\{A_{t-2}A_{t-1}A_t\}$. For $T_1 \neq T_2$ the statistic can be different in that the same history $\{A_{t-2}A_{t-1}A_t\}$ can be followed by the futures A_{t+1} and A'_{t+1} with different probability distributions for T_1 and T_2.

trajectories is explored. Our results demonstrate that the latter is very different for the solvent and the protein atoms.

3. The method

3.1. The observable A_i

The way we construct the observable A for the analysis in the framework of Computational Mechanics involves several steps. Each step is based on well grounded theoretical approaches and verified by extensive numerical tests.

First, we have chosen the velocity of one of the hydrogens as a continuous molecular signal. Then, the velocity was sampled in a way that resembles the construction of the Poincare section thus providing the points on the trajectory at the average intervals of 0.03 ps. This value corresponds to the first minimum on the velocity autocorrelation, a recommended in signal analysis sampling rate.

It turns out that for long enough histories in the Computational Mechanics framework it is sufficient to choose a very coarse-grained observable such as a symbol that can take on a value from a finite alphabet. In other words, the atom's velocity space can be partitioned into few areas labelled with symbols (Fig. C.1). We have tested the cases of 2, 3, 4, and 5 symbols alphabets and found that starting from size 3 alphabet and higher the results produced by Computational Mechanics are essentially the same. Therefore for all our calculations we used symbols from the alphabet {0, 1, 2}.

The described procedure resulted in a sequence of typically two to five millions symbols that was shown to contain most of the statistical information from the original molecular trajectory (see Appendix C for details).

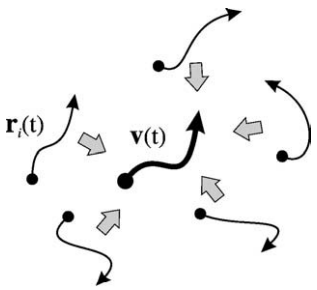


Fig. 3. The atom's velocity time signal $\mathbf{v}(t)$ is a function of the trajectories of the neighbouring atoms $\mathbf{r}_i(t)$ (see text for details).

3.2. The statistic: Computational Mechanics

In the framework of Computational Mechanics the "past" sequences of length l are formed from the analysed signal together with their probabilities (computed as the number of occurrences of a sequence divided by the total number of sequences). The pasts are then grouped into the "causal states" ϵ_i using the criteria of the statistical equivalence of the "futures" following each "past": $\epsilon_i \equiv \{\lambda\colon P(A^+|\lambda) = P(A^+|A_i^-)\}$, where A^- and A^+ are pasts and futures respectively. Each state has its own probability defined by the probabilities of the pasts constituting the state. A matrix of transition probabilities from state to state defines the Markov sequence of the causal states in the signal (note: the signal itself can be non-Markovian). The collection of the causal states together with the transition probabilities constitutes the ϵ-machine. The rigorous definition of the ϵ-machine and its mathematical properties are given in Appendix A.

The *Statistical Complexity* 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: $C_\mu = H[P(\epsilon_i)]$, where P are the probabilities of the states and H is the Shannon entropy of the distribution of a random variable ν, $H[P(\nu)] \equiv -\sum_\nu P(\nu) \log_2 P(\nu)$. ϵ-machines can be reconstructed from observed data using the CSSR algorithm described and implemented in [13].

As an illustrative example, consider an infinite sequence of symbols from the alphabet {0, 1}: …0101010101…. The ϵ-machine for this signal is shown in Fig. 4. Causal state B consists of all pasts of the form …0101, while the state C contains the pasts …1010. When leaving the state B

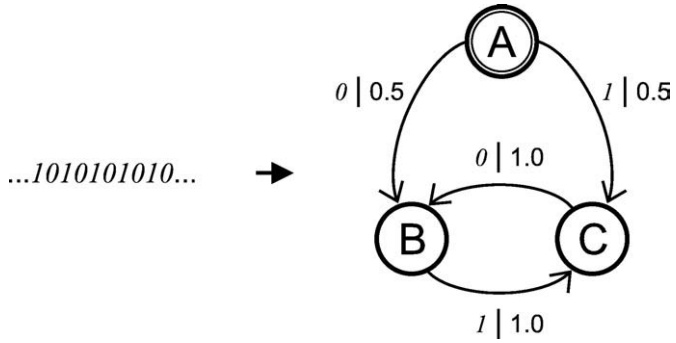


Fig. 4. ϵ-machine for the symbolic sequence shown on the left. *A*, *B*, and *C* are causal states. The numbers on the arrows show the transition probabilities between the states when emitting either 0 or 1.

Author's personal copy

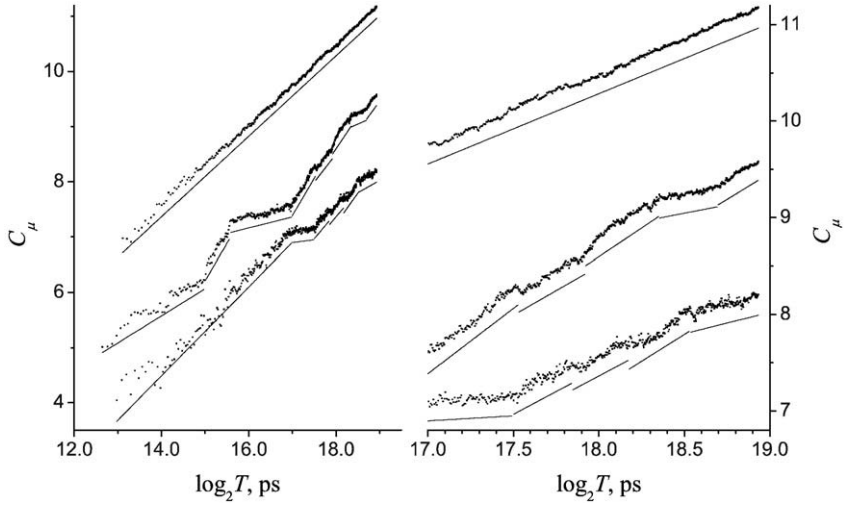


Fig. 5. Dependence of C_μ on the logarithm of the trajectory length T. From top to bottom: the hydrogen of one of the water molecules, the nitrogens of the first and third residues of the peptide respectively. The solid lines represent the linear fits, shifted downwards for clarity.

symbol 0 is emitted with probability 1 and the process is transferred to state C (when adding 0 to the history …0101, that belongs to the state B, the current past becomes …1010). The initial state A, containing both histories …0101 and …1010, is required to start the process and the probabilities of emitting either 0 or 1 on leaving this state are equal.

A finite length of the pasts and futures has to be chosen in practical calculations. We have tested various lengths and found that from the length 6 and longer the results are essentially the same. Therefore, we have used the length 9 in all the calculations.

Thus, the dynamical quantity of interest is the Statistical Complexity C_μ. We are interested in the behaviour of this quantity at different lengths of the signal T' (see Section 2). Therefore, we calculated the values of C_μ as a function of the signal length (simulation time) T.

4. Molecular system and simulation details

We have chosen a 21-residue peptide $A_5(A_3RA)_3A$ from the review [14] where it is reported to fold in 0.8 μs on average. The forcefield for the simulations was GROMOS96 [15]. The peptide was solvated by 1658 SPC water molecules [16] and after proper minimisation of the system's energy was simulated for 0.5 μs using the GROMACS molecular dynamics [17] package. The temperature and the pressure of the system were kept constant at 300 K and 1 bar respectively using Berendsen [18] thermostat. A sufficient equilibration was performed before collecting data for analysis. We have not reached the folded state, however, prolonged periods of the existence of β-sheet and α-helix motifs were recorded (see Section 5). 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 C for the signal processing details).

5. Results

We have found that the Statistical Complexity C_μ grows logarithmically with the signal length T. This is clearly seen in the log_2T-C_μ coordinates, Fig. 5. The data can be fitted by the curve $C_\mu = a + h_Q \log_2 T$ where the parameter h_Q quantifies the growth of the Statistical Complexity value. h_Q characterises the changes in the ϵ-machine, that is the changes in the statistic on the histories (see Section 2). The changes are defined by the rate with which the space of dynamically allowed trajectories of the neighbouring atoms is sampled (explored).

The main result reported in this study is the qualitative and quantitative difference in h_Q for water and protein atoms (Fig. 5). While for water the environmental configurations are explored uniformly, producing a perfect line on the log_2T-C_μ plot, the peptide exhibits well pronounced periods with significantly different rates of the exploration. 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. This can also be seen in Fig. 6 where the same fitted curves were plotted in linear time coordinates.

In the figure the values of h_Q are also plotted for the corresponding periods. For some periods the growth is faster, for others—slower than for water. Sometimes the exploration rate becomes very low, signalling that the peptide atoms environment has covered almost completely the (dynamically) allowed area. The number of possible combinations of the neighbouring atoms trajectories is extremely large (even for only energetically allowed phase space areas) and they can not be exhaustively sampled in the relatively short time of the simulation. Therefore, the slow change in the statistic on the histories indicates a dynamical frustration at these periods of the protein's

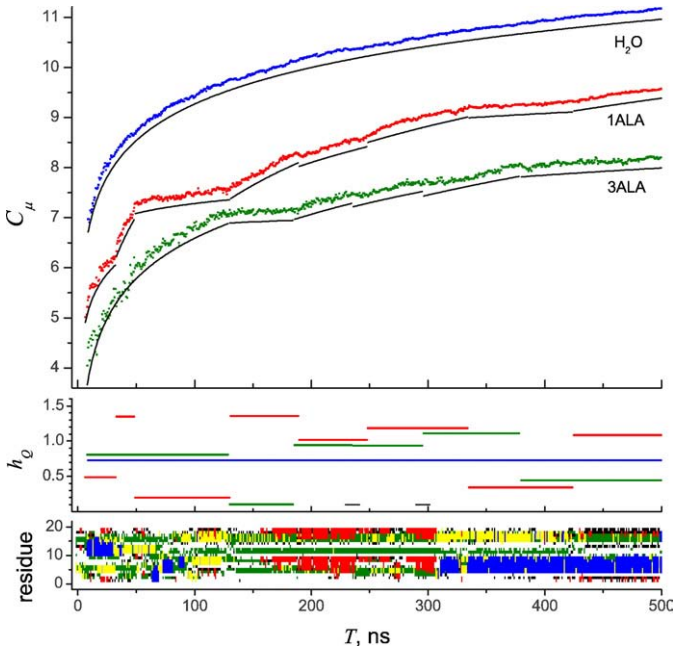


Fig. 6. Top: the trajectory length dependence of C_μ for: blue—water hydrogen, red—residue 1 nitrogen, green—residue 3 nitrogen. Middle: the values of h_Q at the intervals of the constant growth as in Fig. 5. Bottom: the assignment of the peptide's structural motifs for each residue; red—β-sheet, blue—α-helix, yellow—turn, green—bend, black—β-bridge.

Author's personal copy

evolution. In other words, the neighbourhood of the atom moves along a very limited set of routes selected by the dynamics from the space of all possible trajectories. It is interesting to note, however, that on average the value of h_Q does not differ substantially from that of water molecules. This is, perhaps, the indication of the fact that h_Q is a characteristic of the whole dimensional phase space, rather than the dynamics of the individual atoms under consideration. The dynamical frustration found here could be related to the general considerations of the emergence of complex dynamics in systems possessing explicitly this sort of frustration [19].

We have also plotted the classification of the structural motifs of the peptide generated by the DSSP algorithm [20], Fig. 6. The figure illustrates the same time scale of the changes in the folding motifs and the values of h_Q. It is difficult to extract specific correlations between the two and data on other peptide's atoms is needed for more substantiated analysis, which is the subject of our current work.

6. Conclusions

A new measure h_Q has been introduced that characterises the dynamics of molecular system at the nanoseconds time scale. The measure quantifies the way the dynamical patterns in the trajectories of the neighbouring atoms are explored. For water h_Q is found to be constant that implies a uniform covering of the patterns. For the peptide atoms, h_Q exhibits well separated periods of very different rates of the environment exploration. In some periods the values of h_Q are very low suggesting small volumes of the dynamically allowed configurational space, a dynamical frustration. For others, the rate is significantly higher than that of water.

Since the lengths of these periods are of the order of tens of nanoseconds, they can potentially be correlated with the structural changes of the peptide, because the changes belong to the same time scale. Further investigations that would include other atoms of the peptide are required and they are the subject of our current work.

Acknowledgements

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

Computational Mechanics analyses symbolic dynamics. All past A_i^- and future A_i^+ halves of bi-infinite symbolic sequences 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*.

It is proven [21] 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 cannot be subdivided into smaller states;
- a *unique minimal sufficient* statistic, any other one simply re-labels the same states.

The *Statistical Complexity* 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: $C_\mu = H[P(\epsilon_i)]$, where H is the Shannon entropy.

Appendix B. 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 bulk SPC water at 300 K of 30 ns long, that is ≈1 million data points), the alphabet size K and the length l of the histories A^- used by the ϵ-machine reconstruction algorithm CSSR.

The dependence of C_μ on both parameters is shown in Table B.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 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 B.1.

Varying the sampling intervals in converting the velocity signal to discrete times 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, and such partition was used in all subsequent calculations.

The influence of particular MD models and the parameters of the numerical methods on the phenomenon were insignificant. They are as follows.

Table B. 1
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 bulk water hydrogen velocity signal

l	C_μ	K	C_μ
2	3.17	2	5.22
3	4.75	3	7.95
4	6.11	4	8.23
5	7.31	5	8.68
6	7.95		
7	8.15		
8	8.21		
9	8.29		
10	8.37		

- Both Nose-Hoover and Berendsen thermostats produced almost identical results in C_μ with the same $\log_2$-like behaviour. Varying the coupling constant of the Berendsen thermostat by two orders of magnitude did not change the results.
- An SPC-E water model produced slightly higher values of C_μ than SPC while keeping the same overall behaviour of the curves unchanged.
- Systems containing 392 and 878 water molecules resulted in the same values of the complexity parameters.
- Varying the position of the Poincare section plane along the z axes did not lead to any change in the results. The same behaviour with the number of data points was obtained, except that the time between the data points was larger for obvious reasons.
- Finally, different values of the second adjustable parameter of the CSSR algorithm, the significance level for the χ-squared significance test, 0.001, 0.01, and 0.1, reproduced the same qualitative behaviour of C_μ.

Appendix C. Converting molecular trajectory into symbolic sequence

C.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

Author's personal copy

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. C.1.

C.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) [22] 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 $\mathbf{x}_{i+1}=\mathbf{f}(\mathbf{x}_i)$, f: M→*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, $\{A_i\}_{i=-n}^{n}$ where A_i are symbols from the alphabet that mark the elements where $\mathbf{x}_i \epsilon B_i$. 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 [23] 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 [24] 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. C.1. 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.

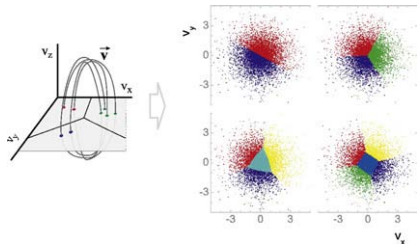


Fig. C.1. Approximations for generating partitions obtained using the method by Buhl and Kennel [24] for the discretised hydrogen velocity for 2, 3, 4, and 5 partitions.

References

[1] S.G. Dastidar, C. Mukhopadhyay, Phys. Rev. E 68 (2) (2003) 021921.
[2] A.R. Bizzarri, S. Cannistraro, Europhys. Lett 37 (3) (1997) 201–206.
[3] C. Rocchi, A.R. Bizzarri, S. Cannistraro, Phys. Rev. E 57 (3) (1998) 3315–3325.
[4] W. Nadler, A.T. Brunger, K. Schulten, M. Karplus, Proc. Natl. Acad. Sci. 84 (22) (1987) 7933–7937 URL http://www.pnas.org/cgi/content/abstract/84/22/7933.
[5] A. Berezhkovsky, G. Sutmann, Phys. Rev. E 65 (2002) 060201.
[6] A. Rahman, Phys. Rev. 136 (1964) 405–411.
[7] M. Mazza, N. Giovambattista, F. Starr, H. Stanley, Phys. Rev. Lett. 96 (2006) 057803.
[8] J.P. Crutchfield, K. Young, Phys. Rev. Lett. 63 (2) (1989) 105–108.
[9] J.P. Crutchfield, K. Young, in: W. Zurek (Ed.), Entropy, Complexity, and Physics of Information, SFI Studies in the Sciences of Complexity, VIII, Addison-Wesley, Reading, Massachusetts, 1990.
[10] J.P. Crutchfield, Physica D 75 (1–3) (1994) 11–54.
[11] D. Nerukh, Chem. Phys. Lett. 457 (4–6) (2008) 439–443;
D. Nerukh, Chem. Phys. Lett. 459 (1–6) (2008) 203.
[12] D. Nerukh, V. Ryabov, R.C. Glen, Phys. Rev. E 77 (2008) 036225.
[13] C.R. Shalizi, K.L. Shalizi, 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.
[14] J. Kubelka, J. Hofrichter, W.A. Eaton, Curr. Opin. Struck. Biol. 14 (1) (2004) 76–88.
[15] W. Scott, P. Hunenberger, I. Tironi, A. Mark, S. Billeter, J. Fennen, A. Torda, T. Huber, P. Kruger, W. van Gunsteren, J. Phys. Chem. A 103 (1999) 3596.
[16] H.J.C. Berendsen, J. Postma, W. van Gunsteren, J. Hermans, in: B. Pullman (Ed.), Intermolecular Forces, D. Reidel Publishing Company, Dordrecht, 1981, pp. 331–342.
[17] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, J. Comp. Chem. 26 (2005) 17011718.
[18] H.J.C. Berendsen, in: M. Meyer, V. Pontikis (Eds.), Computer Simulations in Material Science, Kluwer, 1991, p. 139155.
[19] P.M. Binder, Science 320 (5874) (2008) 322–323 URL http://www.sciencemag.org.
[20] W. Kabsch, C. Sander, Biopolymers 22 (1983) 2577–2637.
[21] C. Shalizi, K. Shalizi, R. Haslinger, Phys. Rev. Lett. 93 (11) (2004) 118701–1-118701-4.
[22] S. Wiggins, Introduction to Applied Nonlinear Dynamical Systems and Chaos, Springer, New York, 1990.
[23] O. Gornerup, K. Lindgren, personal communication (2006).
[24] M. Buhl, M.B. Kennel, Phys. Rev. E 71 (2005) 046213.