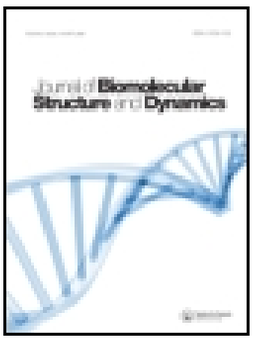

Taylor & Francis
Taylor & Francis Group

Journal of Biomolecular Structure and Dynamics

ISSN: 0739-1102 (Print) 1538-0254 (Online) Journal homepage: https://www.tandfonline.com/loi/tbsd20

Numeric analysis of reversibility of classic movement equations and constructive criteria of estimating quality of molecular dynamic simulations

Ivan Yu. Torshin, Vladimir A. Namiot, Natalia G. Esipova & Vladimir G. Tumanyan

To cite this article: Ivan Yu. Torshin, Vladimir A. Namiot, Natalia G. Esipova & Vladimir G. Tumanyan (2020): Numeric analysis of reversibility of classic movement equations and constructive criteria of estimating quality of molecular dynamic simulations, Journal of Biomolecular Structure and Dynamics, DOI: 10.1080/07391102.2020.1773927

To link to this article: https://doi.org/10.1080/07391102.2020.1773927

JOURNAL OF BIOMOLECULAR STRUCTURE AND DYNAMICS
https://doi.org/10.1080/07391102.2020.1773927

Taylor & Francis
Taylor & Francis Group

Numeric analysis of reversibility of classic movement equations and constructive criteria of estimating quality of molecular dynamic simulations

Ivan Yu. Torshin[a], Vladimir A. Namiot[b], Natalia G. Esipova[c] and Vladimir G. Tumanyan[c]

[a]Dorodnicyn Computing Centre, Russian Academy of Sciences, FIC IU RAS, Moscow, Russian Federation; [b]Institute of Nuclear Physics, Lomonosov Moscow State University, Moscow, Russian Federation; [c]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russian Federation

Communicated by Ramaswamy H. Sarma

ABSTRACT

The fundamental criteria of the quality of molecular dynamics (MD) simulation represent a pivotal challenge, especially in the case of MD simulations of large systems (in particular, proteins).This work presents a simple theoretical analysis of time reversibility in classical mechanics that has allowed us to formulate a number of constructive criteria for evaluating the quality of the trajectories, generated in MD simulations. The results of testing the criteria on the structures of eight small proteins are presented. The criteria can be useful for solving different MD problems, such as: choosing the most appropriate thermostats for a MD system under study, the methods for sampling conformations, etc.

ARTICLE HISTORY
Received 13 April 2020
Accepted 19 May 2020

KEYWORDS
Proteins; molecular dynamics; quality criteria; time reversibility; optimal thermostating

Introduction

Molecular dynamics (MD) is an important tool for modeling the conformation dynamics of proteins and other biomolecules. The MD methods can be used in pharmacophore modeling the potential drugs (Vora et al., 2020), modeling the ligand-receptor interactions (Hu et al., 2018; Wang et al., 2020; Windshügel, 2019; Zhang et al., 2020), simulations of the dynamics of nucleic acid (Pant, Patnak, & Jayaram, 2020), modeling of the solvent effects on the protein structure (Prakash et al., 2018), lipid modeling (Bozdaganyan et al., 2019; Lien et al., 2019), estimating the protein stability (Mohseni et al., 2019; Xiao et al., 2019), etc. MD can be applied as an independent research method or as an auxiliary method (e.g. for structural elucidation in NMR spectroscopy). In the case of absence of any additional experimental information, the necessity of using some high-quality algorithms for MD is especially obvious.

The problem of developing physically meaningful criteria of assessment of the quality/accuracy of MD algorithms is intimately related to another important problem: the difficulty of an adequate estimation of the error accumulation during the numerical integration of the motion equations. The efforts in this direction are quite justified, because MD of protein systems can be presumably used to analyze folding trajectories (Torshin, 2006). It was shown earlier that even small variations in coordinates and impulses during the calculation of MD-trajectories can influence, for instance, the time reversibility. The latter is closely related to the degree of instability of the MD system: the less stable the MD system is, the quicker the irreversibility begins to dominate the MD simulation (Komatsu & Abe, 2004).

Generally speaking, the estimates of the quality of MD trajectories should be obtained on the basis of certain fundamental criteria. The absence of the fundamental criteria for the quality evaluation of MD trajectories is a key problem in the case of protein structures, because for relatively simple physical systems (e.g. a crystal lattice, water molecules, etc.), some additional experimental data can be found that would allow to verify the physical significance of the results obtained during the MD simulation.

In the case of an arbitrary protein structure no such experimental data are usually available. For instance, the results of a highly complex multi-parallel simulation of the small protein BBA5 only allowed to estimate the theoretical folding constant to be within an order of the experimental constant (Torshin, 2006). Thus, some criteria of a fundamental character are required to evaluate the physical, shall we say, correctness of the results of a MD simulation.

Here, we propose that such criteria can be obtained by analyzing the reversibility of changes in the coordinates of atoms during MD simulation of the time reversal. The time reversibility approaches (T(t) symmetry, i.e. the symmetry of the physical laws with respect to the time reversal) are well developed in the modern theoretical physics. It is shown that the formulas of classical mechanics, classical electrodynamics, quantum mechanics, and theory of relativity do not change with the time reversal. T-symmetry is broken in weak interactions in microscopic, as well as macroscopic systems due to the second thermodynamics law and the

CONTACT Ivan Yu. Torshin tiy135@yahoo.com

Supplemental data for this article can be accessed online at https://doi.org/10.1080/07391102.2020.1773927.

degeneracy of macrostates of thermodynamic systems (Nishijima, 1963).

The irreversibility of real thermodynamic systems is associated with the problematics of the Loschmidt paradox, in the framework of which it is argued that a sequence of states with a decreasing entropy is possible for any mechanical system due to time reversibility of the Newtonian dynamics equations. At the same time the phase trajectories of multiparticle systems are obviously stochastic, that eliminates contradiction between the reversibility of the classical mechanics laws and the irreversibility of thermodynamic macrosystems (Karlov & Kirichenko, 2011). Time-reversibility in MD was previously used for studying the dynamics of classical fluids (Levesque & Verlet, 1993).

Being on the edge between macro- and microsystems, proteins represent a very interesting case for MD simulations. On one hand a protein molecule is a system of many particles. On the other hand a protein molecule can be considered as one massive particle (since in most cases protein chain, formed by covalent bonds, is held together by hydrophobic and electrostatic forces). Therefore some fundamental criteria of the accuracy of the MD algorithms are of extreme importance for modeling the protein conformation.

On the time reversibility and the constructive criteria for evaluating the quality of MD simulation procedures

Despite the phenomenon of the so-called spin echo (Tanner & Stejskal, 1968), which can be interpreted as the time reversibility, studying the time reversibility at atomic level is virtually impossible in proteins. Nevertheless, it is quite possible to model the real molecular systems by numerical methods based on classical mechanics (i.e. MD simulation). Obviously, if a protein molecule is being simulated within a numerical procedure, some changes to the values and signs of any parameter of the simulated particle system can be introduced at any time during the simulation. This technical opportunity draws attention to the time reversibility as the basis for developing criteria for evaluating the quality of the MD simulation of the complex systems.

Let us take a system C consisting of N particles, $C = \{(q_i, p_i)\}, i = 1 \ldots 3N$, where q_i is the coordinate, and p_i is momentum of the *i*-th particle. The state of the system C at time moment t is denoted as $C_t = \{(q_i(t), p_i(t))\}$. The Hamilton equations for the system C allow us to calculate the derivatives of coordinates and momentums [18] (Landau & Lifshitz, 1976):

$$\dot{q}_i = \frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}. \dot{p}_i = \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}. \quad (1)$$

The value of the coordinate q_i at an arbitrary time point t is calculated in accordance with a recurrent formula such as

$$q_i(t) = q_i(t-dt) + dq_i, dq_i = \frac{\partial H}{\partial p_i} dt, q_i(0) = q_{i,0}. \quad (2)$$

Definition 1. **R-transform.** The time reversibility transformation R in classical mechanics is determined by the following set of rules (Nishijima, 1963): 1) $q_i^R = q_i$, $p_i^R = -p_i$; 2) physical quantities are not changed, if they are not dynamic variables (mass, charge, etc.); 3) $H^R = H$.

Hamilton equations (1) and the R-transform allow us to calculate the particle coordinates at any time moment after the time reversal. The i-th coordinate value, calculated after the time reversal at the time moment τ_x, is denoted as $q_i^{'\tau_x}(t)$, and the state of the entire system of particles is denoted as $C_t^{'\tau_x}$ or simply C_t'.

Theorem 1. *Carry out the following imaginary experiment with system of N particles,* $C = \{(q_i, p_i)\}, i = 1 \ldots 3N$. *Let the system evolve from the time* $t = 0$ *till an arbitrary time* $t = \tau_x$ *in accordance with equations (1, 2). R-transformation is applied at* $t = \tau_x$. *Subject to the energy conservation law* $q_i^{'\tau_x}(t) - q_i(t) = 0$ *for all* $i = 1 \ldots 3N$ *and all* $t \le \tau_x$.

Proof. At the time point τ_x, the coordinates of the i-th particle are calculated as $q_i(\tau_x) = q_i(\tau_x - dt) + dq_i$. According to Definition 1, the Hamilton equation for $\dot{p}_i$ is not changed with the time reversal, the Hamiltonian is not changed too, and in the Hamilton equation for the coordinates $\dot{q}_i$ the sign is changed (as $p_i^R = -p_i$). Applying the R-transform at the time τ_x, we obtain the equation $dq_i = -\frac{\partial H}{\partial p_i} dt$, and can calculate the value of an arbitrary coordinate q_i at the immediately preceding time moment τ_x, $\tau_x - dt$. If the energy conservation law is fulfilled, then $\frac{dH}{dt} = 0$, so the expression for dq_i can be substituted in recurrent formula (2) to obtain $q_i^{'\tau_x}(\tau_x - dt) = q_i^{'\tau_x}(\tau_x) - \frac{\partial H}{\partial p_i} dt$ (it is obvious that $q_i^{'\tau_x}(\tau_x) = q_i(\tau_x)$).

Ar the same time, $q_i(\tau_x) = q_i(\tau_x - dt) + dq_i$. Therefore, $q_i^{'\tau_x}(\tau_x - dt) = q_i(\tau_x - dt)$. Correspondingly for all $0 \le t \le \tau_x$ $q_i^{'\tau_x}(t) = q_i(t)$, i.e. $q_i^{'\tau_x}(t) - q_i(t) = 0$. Thus, the classical system has to repeat exactly its traversed path after the time reversal transformation, but in the opposite direction. The theorem is proved.

Corollary 1. *For the system* C *being reversed at the time moment* τ_x, *we introduce the quantity* $rmsd(C_0, C_0^{'\tau_x}) = \sqrt{\frac{1}{3N}\sum_{i=1}^{3N} (q_i^{'\tau_x}(0) - q_{i,0})^2}$. *Under the theorem conditions* $rmsd(C, \tau_x) = 0$.

Corollary 2. *Let* $rmsd(C_0, C_{\tau_x}) > 0$ *for all* $\tau_x > 0$. *Define the quantity* $\kappa(\tau_x, C) = rmsd(C_0, C_0^{'\tau_x}) / rmsd(C_0, C_{\tau_x})$. *Under the theorem conditions* $\kappa(\tau_x) = 0$.

Corollary 3. *Let* $\sigma(C_0, C_0^{'\tau_x}) = \frac{1}{3N}\sum_{i=1}^{3N} \int_0^{\tau_x} (q_i^{'\tau_x}(t) - q_i(t))^2 dt$. *Under the theorem conditions* $\sigma(C_0, C_0^{'\tau_x}) = 0$.

Theorem 1 and its Corollaries can be considered as a fundamental basis for introducing a number of constructive criteria for the quality evaluation of the MD simulation procedures. Consider the discrete case of integration of the motion equations when $q_i(t) = q_i(t - \Delta t) + dq_i + \varepsilon_i(t)$, where $\Delta t > 0$ is a time step of the MD simulation (typically it is 1 fs or less), $\varepsilon_i(t) > 0$ is an integration error that depends on time (as integration errors are accumulated at each step) and $\varepsilon_i(0) = 0$. The error $\varepsilon_i^{'\tau_x}(t)$, $0 \le t \le \tau_x$ corresponds to the calculation of the coordinates after the time reversal at a time τ_x. Hence,

Table 1. Proteins under study.

Protein	PDB	N_{aa}	N_{at}	Comments
Fs-peptide	–	19	234	$(Ala)_7$-Arg-$(Ala)_4$-Arg-$(Ala)_4$-Arg-Ala
FsA-peptide	–	19	192	Polyalanine helix
TC5B	1L2Y	20	303	Synthetic protein «TRP-CAGE»
BBA5	1T8J	23	378	Synthetic protein, the folding model
Protein G	2GB1	56	854	Immunoglobulin-binding protein S.Aureus
BPTI	1G6X	58	893	Pancreatic Trypsin Inhibitor
Protein B	1BDD	60	940	Immunoglobulin-binding protein S.Aureus
A3D	2A3D	73	1139	Synthetic protein, 3 helices

Notes: Column N_{aa} is a number of amino-acid residues; N_{at} is a number of atoms (including hydrogen atoms).

for all $\tau_x>0$ for the system under study, the value e.g. of the functional rmsd(C, τ_x) (Corollary 1) is not certainly be equal to 0, since $\mathrm{rmsd}(C, \tau_x) = \sqrt{\frac{1}{3N}\sum_{i=1}^{3N}(\varepsilon_i^{'\tau_x}(0))^2}$.

Thus, the functionals $\mathrm{rmsd}(C_0, C_0^{'\tau_x})$, $\kappa(\tau_x, C)$ and $\sigma(C_0, C_0^{'\tau_x})$ allow evaluating the «accuracy» of the MD trajectory, passed by the system from $t=0$ to $t=\tau_x$ and backward to the time point $t=0$ after the time revesal. At the same time the values of these functionals allow us an indirect evaluation of the integration error of the motion equations, that is averaged over the coordinates of all particles of the system C.

Materials and methods

Molecular dynamics system in this article is a combination of: (A) the calculation method (a set of algorithms implemented as a program), (B) the parameters of the calculation method (including initial conditions for MD simulations) and (C) the simulated object (a system of particles with their initial coordinates).

The *calculation method* was implemented in the ECMMS software system (Torshin, 2004) and includes the following computational components: 1) calculating the potential energy of a system, 2) minimizing the energy of the system (that is essential technically to start a stable MD simulation), 3) selecting the initial conditions (velocities, since the coordinates of atoms are determined experimentally), 4) calculating the coordinates at the next time moment, 5) modeling the solvent (in the presented series of experiments in vacuo, i.e. without introducing the special term and for $\varepsilon=1$), and 6) controlling the temperature (a procedure of imposing limitations and corrections for admissible values of particle velocities or velocity distrubutions).

The *studied objects* (protein molecules) are listed in Table 1, the structures of the initial conformations are shown in Supplementary Figure 1.

The potential energy U_{pot} of the molecular system containing N atoms, B covalent bonds (binding terms), A angular terms and T torsion terms was calculated as $U_{pot} = \sum_{j=1}^{B} E_{bond}^j + \sum_{j=1}^{A} E_{ang}^j + \sum_{j=1}^{T} E_{tors}^j + \sum_{j=1}^{N}\sum_{k=j+1}^{N} E_{nonb}^{j,k}$, where $E_{bond}^j = \frac{1}{2}K_b^j(r^j - r_0^j)^2$ (the energy term of the j-th covalent bond), $E_{ang}^j = \frac{1}{2}K_\theta^j(\theta^j - \theta_0^j)^2$ (the deformation energy term of the j-th valent angle between two covalent bonds), $E_{tors}^j = \frac{1}{2}\sum_{m=1}^{n} V_m^j(1+\cos(m\phi^j - \phi_0^j))$ (the rotation energy term of the j-th dihedral angle). The term of non-bonding interactions was defined as the sum of the electrostatic and Van-der-Waals terms $E_{nonb}^{j,k} = E_{el}^{j,k} + E_{vdW}^{j,k}$. The electrostatic interactions were described by Coulomb's law $E_{el}^{j,k} = q_j q_k/4\pi\varepsilon\varepsilon_0 r_{j,k}$, and the Van-der-Waals interactions are described by the Lennard-Jones potential $E_{vdW}^{j,k} = -\frac{B_{jk}}{r_{jk}^6} + \frac{A_{jk}}{r_{jk}^{12}}$.

The force constants K_b, K_θ, V_m, the equilibrium values of the bonds and angles r_0, θ_0, φ_0, as well as the constants A_{jk}, B_{jk} were defined in accordance with the UFF potential [19] (Rappe et al., 1992). The values of the partial atomic charges q_j, q_k were determined on the basis of the AMBER potential [20] (Weiner et al., 1986) based on quantum-mechanical calculations. This «UFF + AMBER» potential was implemented earlier within the framework of the ECMMS software system (Torshin, 2004).

The **potential energy was minimized** by the method of conjugated gradients (when displacement of each atom occurs in the direction of forces acting on the atom). For a molecule of N atoms, the force acting on every atom was defined as $\vec{F}(\vec{x}_j) = \vec{\nabla}_{x_j} U_{pot}$, $j=1.N$. For the i-th step of the minimization by conjugated gradients, the new coordinates of every atom were calculated as $\vec{x}_j^i = \vec{x}_j^{i-1} + k\cdot\vec{h}_j^i$, where $\vec{h}_j^i = \vec{F}(x_j^i) + \frac{|\vec{F}(\vec{x}_j^i)|^2}{|\vec{F}(\vec{x}_j^{i-1})|^2}\vec{h}_j^{i-1}$

Molecular dynamics was based on the second Newton's law $\vec{F}(\vec{x}_j) = m_j a_j = m_j \dot{V}_j(t)$, $V_j(t) = \dot{x}_j(t)$. The acceleration values a_j of individual atoms were calculated from the sum of the potential and kinetic energies, and then they were numerically integrated using the Verlet's «leapfrog» algorithm (Verlet, 1967). That is, the velocities and coordinates at the time point $t+\Delta t$ were calculated as $\vec{x}(t+\Delta t) = \vec{x}(t) + \vec{v}(t)\Delta t + 0.5\vec{a}(t)(\Delta t)^2$, $\vec{v}(t+\Delta t) = \vec{v}(t) + 0.5\Delta t(\vec{a}(t) + \vec{a}(t+\Delta t))$. **The initial atom velocities** were randomized according to the Maxwell distribution $D(v) = (m/2\pi kT)^{3/2}4\pi v^2 e^{-mv^2/2kT}$. For the calculations $T=300$ K and $\varepsilon=1$ are used, the step was $\Delta t=1$ fs.

Thermostat procedures were implemented by various methods (Table 2). Only deterministic methods for the temperature control were used (Schlick, 2002). Andersen's method was not used since it involves a randomization procedure. For the same reason the stochastic term of the Langevin dynamics was set to zero.

During the **numerical analysis of the obtained criteria** satisfiability a certain simulation time τ_x was chosen. For the initial minimized system (state «0» with a set of coordinates C_0) the MD system calculations were carried out for the given initial set of velocities during the time interval $0\ldots\tau_x$, and the set of coordinates $C_{\tau x}$ was obtained as the result. Then the signs of the velocity components were reversed, and the MD simulation was carried out during the time τ_x «in the opposite direction» in order to obtain the coordinate set $C_0^{'\tau_x}$. The standard (mean-root-square) deviations between the coordinate sets C_0 and $C\tau_x$, rmsd(C_0,$C_{\tau x}$), and between the coordinate sets C_0 и $C_0^{'\tau_x}$, rmsd(C_0, $C_0^{'\tau_x}$) were calculated (Corollary 1 of Theorem 1).

Obviously, if the theorem condition is fulfilled, then rmsd(C_0, $C_0^{'\tau_x}$)=0 for any τ_x, and accordingly for any rmsd(C_0,$C_{\tau x}$). However, the existing MD systems are

Table 2. The studied temperature control models (thermostats).

Model	Description
NONE	No thermostat
VNORM3RT	The scaling of atom velocities is performed in a way that the kinetic energy is 3/2NKT at every step
TNORM3RT	Dynamic velocity scaling (the time step is recalculated to keep the temperature corresponding to the kinetic energy of 3/2NKT)
CMAXWELL	Velocities are recalculated to match the Maxwell distribution corresponding to the given temperature
BERENDSEN $\tau = 50$ fs	Berendsen thermostat with different time constants τ characterizing the thermostat «reactivity». The coefficient λ, by which the
BERENDSEN $\tau = 100$ fs	velocities are multiplied, is calculated as
BERENDSEN $\tau = 300$ fs	$\lambda^2 = 1 + \frac{\delta t}{\tau}\left(\frac{T}{T_{inst}} - 1\right)$
BERENDSEN $\tau = 400$ fs	where T_{inst} is a «temperature» estimated from the velocities at the given time point)
LANGEVIN $\gamma = 0.05$	Langevin dynamics without a stochastic component (as it includes randomization of velocities and hence excludes the time
LANGEVIN $\gamma = 0.30$	reversibility), γ is the coefficient of the 'viscous' term.

imperfect in respect of the methods of potential energy calculation, integration, modeling a solvent, and temperature control, that introduces a certain error $\varepsilon_i(t)$ at every step of the numerical integration and can lead to accumulating a significant total error.

Therefore, to evaluate the satisfiability of the obtained criteria a *return coefficient* $\kappa(\tau_x)$ was introduced, that was calculated as the ratio of rmsd(C_0, $C_0^{'\tau_x}$) to rmsd(C_0,$C_{\tau x}$) (Corollary 2 of Theorem 1). The value $\kappa(\tau_x) = 0$ corresponds to the fulfillment of the obtained reversibility criteria, and the values $\kappa(\tau_x) > 1.0$ correspond to significant violations of the condition of the Theorem 1.

We also introduce *maximum conformational memory time*, $\tau_{max}(\tau_x)$, defined as the maximum simulation time τ_x at which the value $\kappa(\tau_x)$ is not higher than 1.0. One can say that the indicators $\kappa(\tau_x)$ and $\tau_{max}(\tau_x)$ characterize some property of the «reversibility memory» for the studied molecular dynamics system (i.e. the set of the calculation method and the object) in the time interval τ_x.

In the present study the satisfiability of the obtained criteria was investigated for values of τ_x equal to 100, 1000, 5000 and 10000 femtoseconds. For each protein, the calculations were performed for 10 different sets of initial velocities, and the empirical distribution functions of the indicators rmsd(C_0, C_0'), rmsd(C_0,$C_{\tau x}$) and $\kappa(\tau_x)$ were found for every molecular system. The curves were approximatied by the monotonic cubic spline in the coordinates (τ_x, $\kappa(\tau_x)$) and then the values of $\tau_{max}(\tau_x)$ were found as the abscissa of the intersection point of the spline with the graph of the function $\kappa(\tau_x)$=1.0.

The results of numerical modeling the dynamics of protein conformations

Consider in the begining three important conclusions from the results of preliminary experiments related to (1) the study of effects of the individual potential terms within the expression $U_{pot} = E_{bond} + E_{ang} + E_{tors} + E_{nonb}$, and (2) the choice of the integration method.

First, at analyzing the properties of partial potentials (E_{bond}, $E_{bond} + E_{ang}$, $E_{bond} + E_{ang} + E_{tors}$) it has become obvious that the term $E_{nonb} = E_{el} + E_{vdW}$ made the greatest contribution to the sharp decrease of the parameter value $\tau_{max}(\tau_x)$. When the term E_{nonb} was removed from U_{pot}, the value of $\tau_{max}(\tau_x)$ increased by an order of magnitude at least. This conclusion indicates the necessity to select and adjust the parameters of the term E_{nonb} most carefully. The terms E_{el} and E_{vdW} had a comparable effect on reducing the value of $\tau_{max}(\tau_x)$ at that.

Secondly, the integration error of the MD trajectory resulting from restrictions on the storage of real numbers in the computer's memory (word size) was evident only at rather great values of τ_x (tens of nanoseconds, i.e. millions of integration steps over the Verlet algorithm). For example, replacing all real variables of the «double» type (that allows the values in the range from $5{\cdot}10^{-324}$ to $1.7{\cdot}10^{308}$, 8 bytes to store a variable) by «extended» type variables (range $1.9{\cdot}10^{-4932}$ to $1.1{\cdot}10^{4932}$, 10 bytes to store a variable) the noticeable differences in the MD trajectories for the BBA5 protein (rmsd $>$ 0.1 Å) with the same initial conditions were accumulated only after 19 ns of the simulation time.

Thirdly, the complication of the integration scheme (for example, the use of Beeman's algorithm (Beeman, 1976) for integration and more complex schemes) did not result in a significant improvement of the values of rmsd(C_0, $C_0^{'\tau_x}$) and $\kappa(\tau_x)$, $\tau_{max}(\tau_x)$ for the studied systems. The values of these functionals did actually improve (by 10–20% for the Beeman algorithm as compared to the Verlet algorithm) but at the expense of a significant increase of the simulation time (50% at the very least). Therefore, the Verlet integration was used for further experiments.

Consider the results for polyalanine helix (FsA peptide) in more detail. This model system is quite interesting as it contains the same side chains at each position and these side chains are rather short (the methyl group –CH_3). During the simulation of the average duration ($\tau_x = 1000$ fs) the system deviated by an average of 2.23 ± 0.15 Å from the initial C_0 conformation. After reversing the velocities the system evolved into the conformation $C_0^{'\tau_x}$ that was quite close to the initial point C_0 (rmsd(C_0, $C_0^{'\tau_x}$) = 0.34 ± 0.10 Å, averaging over 10 trajectories, Figure 1A), the return coefficient was $\kappa(\tau_x) = 0.15 \pm 0.04$. Thus, this system demonstrates a noticeable «conformational memory» in the studied time interval.

Consider the effects of various thermostats in the same example of the FsA peptide. The analysis of the distributions of the $\kappa(\tau_x)$ values for the FsA peptide (Figure 1B) and other studied proteins (Table 1) has showed, that Langevin dynamics (even if the stochastic component was zero) leads to significant violations of all the obtained criteria compared to all other used thermostats. The similar picture was observed for all of the protein structures: the Langevin dynamics with $\tau_x > 200$ fs resulted in $\kappa(\tau_x)$ values, that were significantly higher than 1.0.

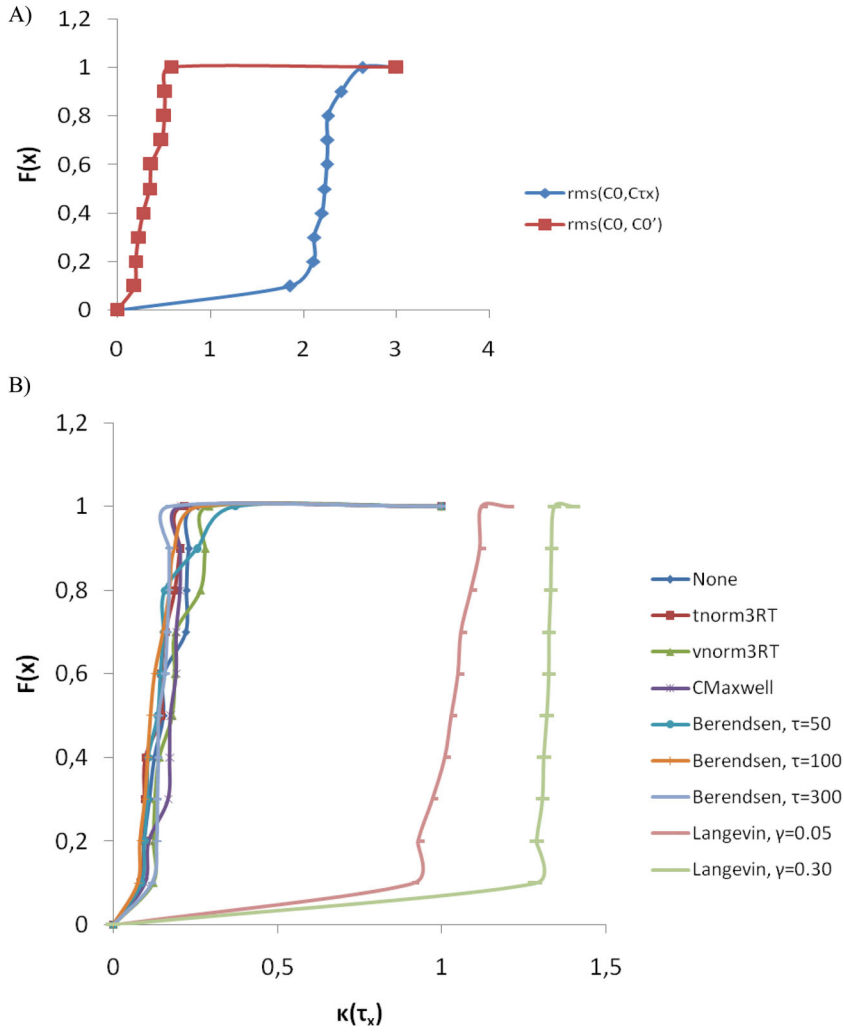

Figure 1. The MD simulation results of the FsA peptide. (A) The empirical distribution function of the r.m.s. deviation of the final structure C^{τ_x} and the «return» structure C_0' from the initial structure (C_0), calculated at 10 sets of initial velocities (τ_x=1000 fs, no thermostat). The value of $\kappa(\tau_x)$ was 0.15 ± 0.04. (B) The effect of various thermostats on the distribution of $\kappa(\tau_x)$ values, τ_x=1000 fs. F(x) is the value of the cumulative empirical distribution function of the indicator $\kappa(\tau_x)$. The worst cases were apparently the thermostats based on Langevin dynamics modeling.

The $\kappa(\tau_x)$ values for the studied eight proteins for $\tau_x = 100$, 1000, 5000, 10000 fs are shown in Supplementary Tables 1–4. The $\kappa(\tau_x)$ values were rather low for $\tau_x = 100$ fs (0.01.0.12). However the $\kappa(\tau_x)$ values did grow with the increasing τ_x. At $\tau_x = 10000$ fs, the $\kappa(\tau_x)$ values for all studied molecular systems were close to 1.0 (except for the polyalanine peptide FsA).

Note that the example with the polyalanine peptide is of particular interest for our studies, as such objects as the conformationally stable peptides are often enriched with alanine (Batyanovskii et al., 2015). Conformationally stable peptides (Batyanovskii et al., 2015; Batyanovskii & Vlasov, 2008; Torshin et al., 2019) are short fragments of the polypeptide chain that tend to hold a certain (unique) conformation coupled to a certain amino acid sequence. The conformationally stable peptides are most often α-helical (Batyanovskii et al., 2015; Torshin et al., 2014; Uroshlev et al., 2015; 2019).

The numerical study of satisfiability of the criteria obtained in Theorem 1 also allows us to evaluate the applicability of various thermostat models to a particular protein. An average over the studied proteins, the smallest $\kappa(\tau_x)$ values were observed without using any thermostat. Among the thermostat models, the smallest $\kappa(\tau_x)$ were observed with Berendsen's thermostat.

The diagrams in coordinates ($\mathrm{rmsd}(C_0,C_{\tau x})$, $\kappa(\tau_x)$), shown in Figure 2, allow better understanding of the satisfiability of the criteria for various τ_x. These diagrams represent the results of all the experiments carried out for the given τ_x. Every point of the diagram is the result of testing one protein with one set of the initial velocities for the given thermostat. Obviously, the position of each point characterizes simultaneously the satisfiability of the obtained criteria (the $\kappa(\tau_x)$ value) and the average amplitude of the molecule conformation changes during the $0 \ldots \tau_x$ run of the MD simulation ($\mathrm{rmsd}(C_0,C_{\tau x})$ value).

The separable clusters of points corresponding to different thermostat models were obviously present at $\tau_x = 100$ fs and at $\tau_x = 1000$ fs. As notice earlier, the Langevin dynamics characterized by a low reversibility ($\kappa(\tau_x) \sim 1$) and, at the same time, by a low amplitude of the conformation changes ($\sim$1 angstrom). At $\tau_x = 5000$ fs when the most part of the systems reach the maximum conformational memory time

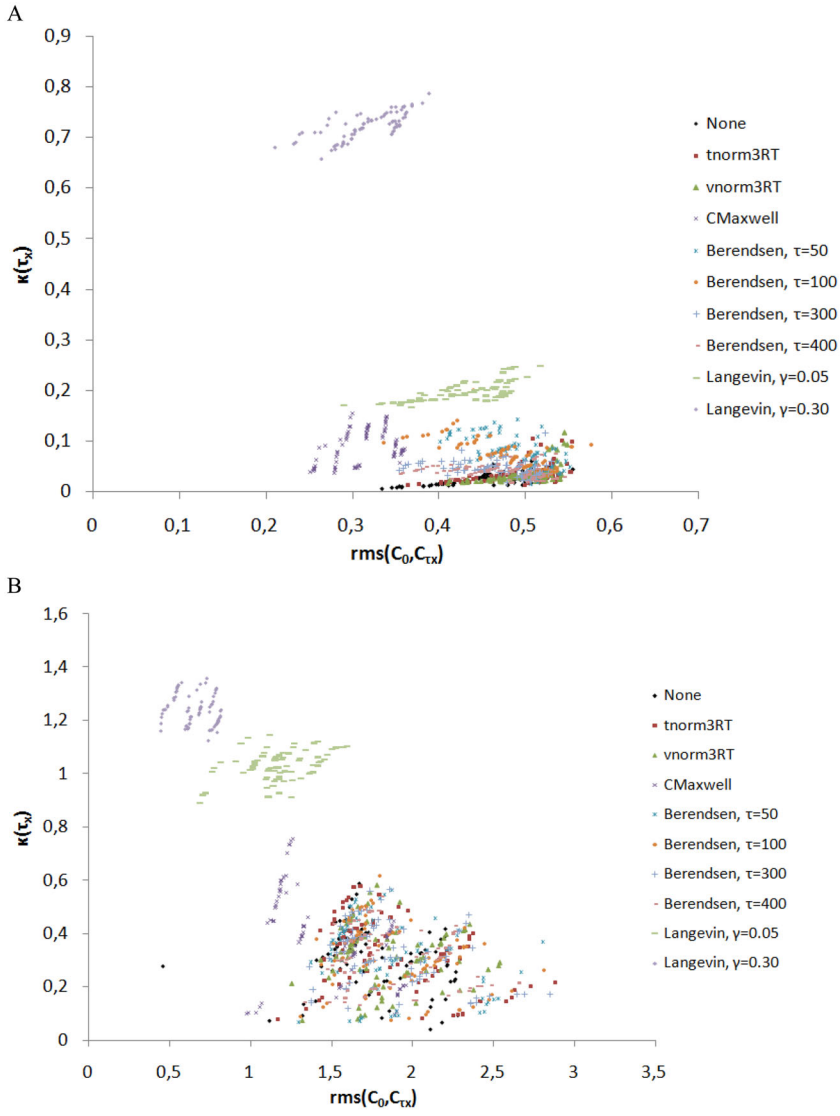

Figure 2. Diagrams in the coordinates (rmsd(C_0,$C_{\tau x}$), $\kappa(\tau_x)$) that present all the carried out experiments at the given value of τ_x. A) τ_x=100 fs, B) τ_x=1000 fs, C) τ_x=5000 fs, D) Several examples, $\tau_x = 5000$ fs.

$\tau_{max}(\tau_x)$, the edges between most of clusters almost disappeared (see also Supplementary Table 5).

The analysis of the diagrams in coordinates (rmsd(C_0,$C_{\tau x}$), $\kappa(\tau_x)$) allowed us to compare the effect of various thermostats on the satisfiability of the obtained criteria. Count, for instance, the number of points (i.e. individual experiments) satisfying the condition ($\kappa(\tau_x)<1$) for every thermostating method and the corresponding percentage (Figure 3). The calculation results for $\tau_x = 5000$ fs show, that the highest percentage (69%) has been observed in the case of Berendsen thermostat with $\tau = 100$. This result was statistically significant by the $\chi 2$ criterion ($p = 0.025$) in comparison with the lowest percentage (51% for «VNORM3RT», i.e. scaling the atomic velocities to match 3/2NKT).

Thus, the study of the satisfiability of the formulated in Theorem 1 criteria allows us to obtain quantitative estimates of some conformational memory of the modeled molecular system (including thermostat, force field parameters, initial coordinate sets etc). Larger values of the return coefficient $\kappa(\tau_x)$ and simulation time τ_x correspond to a smaller degree of conformational memory and, at the same time, to a greater degree of sampling the system's conformations near the equilibrium point. The more complete sampling of the conformations is a significant problem in molecular dynamics (Torshin, 2006). Therefore, the Langevin dynamics represents a more effective method for sampling conformations even with a zero stochastic component, than the other studied thermostat models.

The energy conservation law and the satisfiability of the obtained criteria

The criteria formulated in the Theorem 1 were obtained under the condition of the energy conservation law. The procedure of calculating the MD trajectory of an individual molecule involves some fluctuations of the kinetic and potential energy along the trajectory, which should not go beyond the corresponding distribution. Therefore, in the case of MD the energy conservation law would be more appropriate to consider as an existence of a certain fixed distribution of total energy values (or an admissible interval of the total

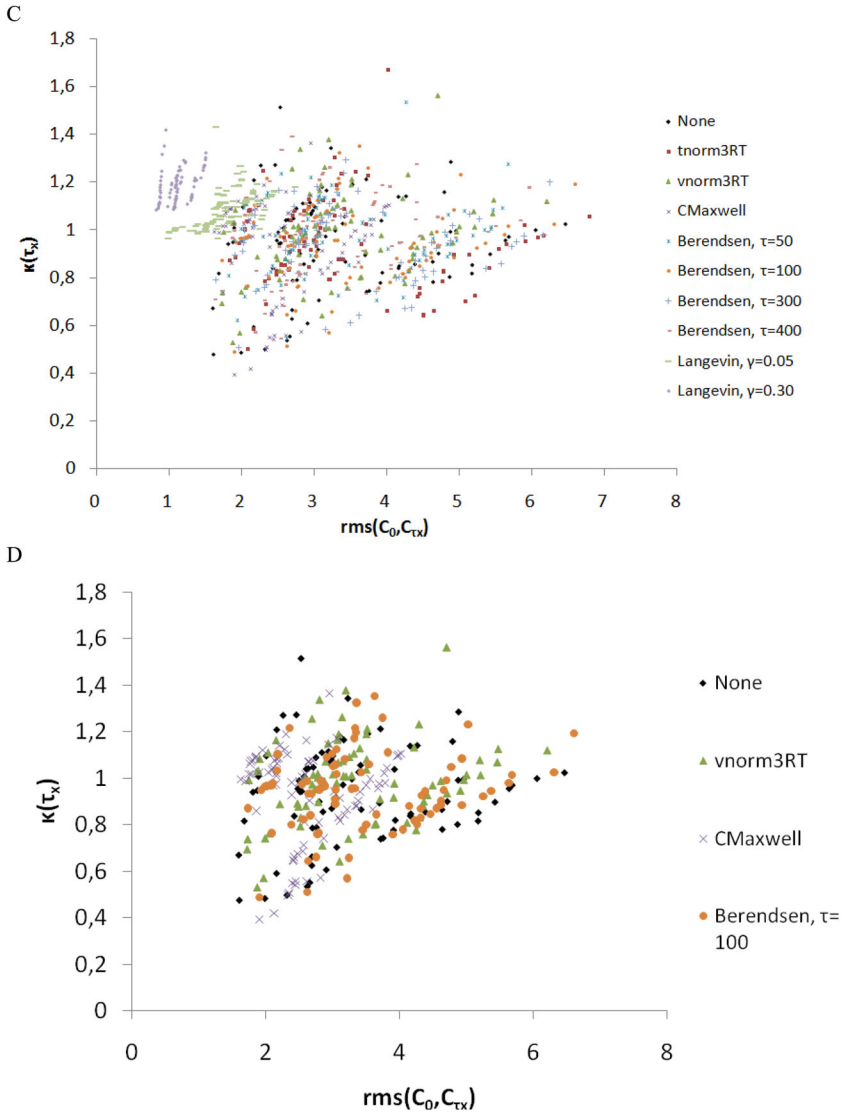

Figure 2. Continued.

energy values) of the molecular system modeled. Then, the reverse trajectory, calculated during testing the satisfiability of the obtained criteria, should not go beyond such an interval.

Below, we consider the results of the energy calculation during direct/reverse trajectories (on the example of the model protein BBA5). The total energy along trajectory for a fixed set of initial velocities is shown in Figure 4A. The energy values should coincide, if the obtained criteria are satisfied fullfilly, but it does not happen due to the errors (integration errors etc.) discussed earlier. Nevertheless, the similarity elements of the direct and the reverse trajectories are evident (in particular, similar localization of the extrema).

The energy calculation results of BBA5 protein for various sets of initial velocities are presented in Figure 4B, τ_x=1000 fs. Obviously, there is a certain range of values in which the values of the total energy (E) fall. It is interesting to note that at τ_x = 1000 fs, the average of the distribution of the E values for the reverse trajectory (3995 ± 91 kJ/mol) was slightly lower than for the direct trajectory (4009 ± 88 kJ/mol, p = 0.00039 according to the Student's test). However, these differences disappeared at $\tau_x \geq$5000 fs. For example, for τ_x = 10000 fs, the corresponding values for the reverse and direct trajectories were 4048 ± 137 kJ/mol and 4047 ± 107 kJ/mol, respectively (p = 0.41).

The average values of the total energy E along the direct and the reverse trajectories, as well as the standard deviations of the E values for all the studied systems are presented in Supplementary Tables 6,7. The linear approximation showed that, the standard deviations of E were directly proportional to the E values (r = 0.8 … 0.9) and were equal to 0.016 … 0.022 for all of the thermostats except the Langevin dynamics.

These results allow us to offer a new approach for thermostating the MD systems, based on the energy conservation law. Namely, it would be necessary to fix not the temperature (that is a thermodynamic parameter in general), but the total energy of the system (that is fully compatible with the classical mechanics at the microscopic, non-thermodynamic level). This energy fixation can be carried out by setting a certain distribution of admissible values of the total energy or by using one fixed energy value.

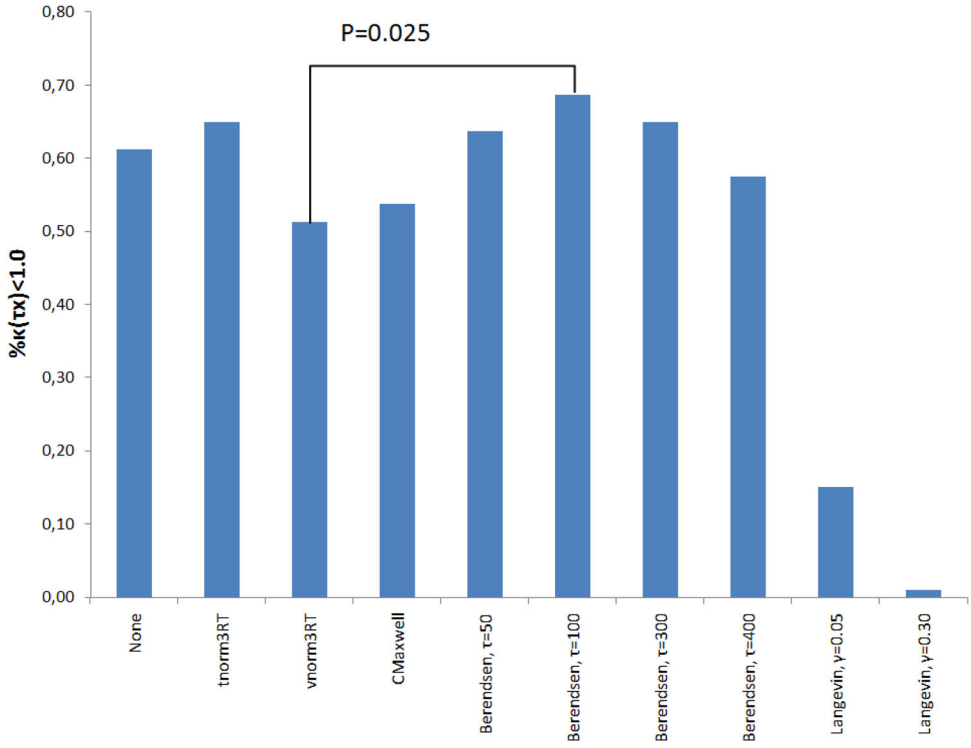

Figure 3. The percentage of individual experiments with the satisfied condition ($\kappa(\tau_x)<1$), τ_x=5000 fs.

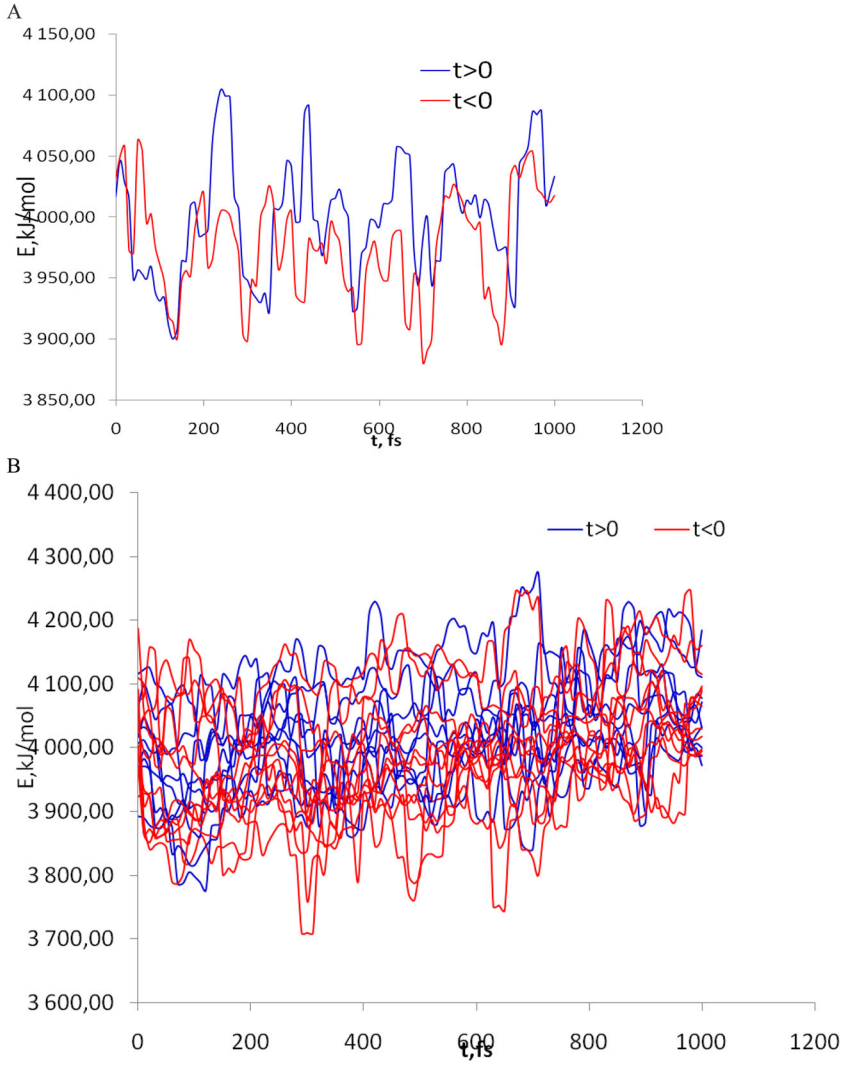

Figure 4. The total energy values for the direct ($t > 0$) and reverse ($t < 0$) BBA5 protein MD trajectories (τ_x = 1000 fs, without any thermostat) A) a fixed set of initial velocities, B) various sets of initial velocities.

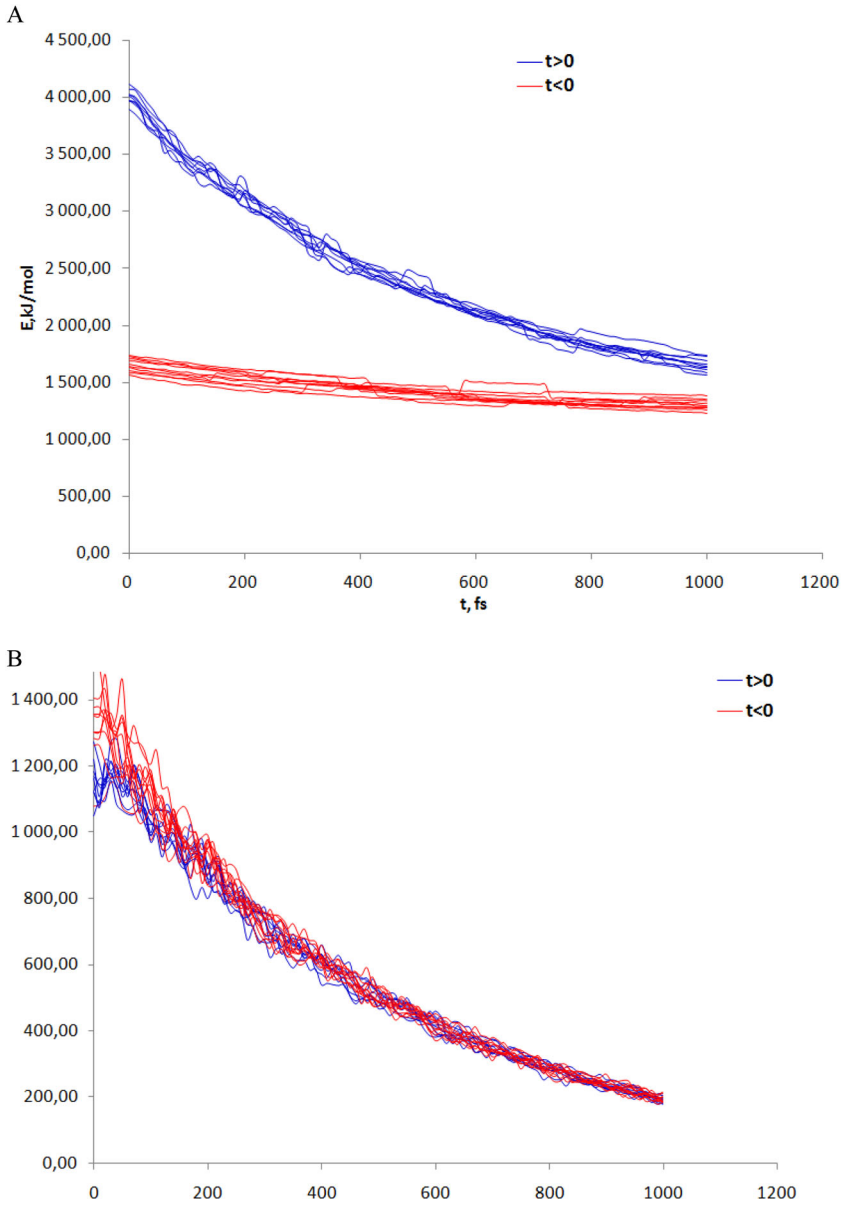

Figure 5. The total energy values for the direct ("$t > 0$") and reverse ("$t < 0$") trajectories of the BBA5 protein (Langevin dynamics, $\gamma = 0.05$, $\tau_x = 1000$ fs, zero stochastic term) A) standard Langevin dynamics, B) Langevin dynamics with the inverse sign of the viscous term (see the text).

The total energy during the Langevin dynamics with the zero stochastic term

It is worth noting that the more complete sampling of the conformations, characteristic for the Langevin dynamics, is connected with systematic changes of the total energy of the simulated systems. Apparently, the introduction of the viscosity term proportional to the particle velocity leads to the energy dissipation of the system (which ceases to be Hamiltonian in this case). Therefore, it is not surprisingly that the worst reversibility of the system along the MD trajectory, estimated by the $\kappa(\tau_x)$ and $\tau_{max}(\tau_x)$ values, was observed just for the Langevin thermostat. The total system energy dissipates and reversing the trajectory with the zero stochastic term would not lead to an energy gain during testing the satisfiability of the obtained criteria. This effect was indeed observed in numerical experiments (Figure 5A).

When the trajectory of the Langevin dynamics is reversed, the sign of the viscous term can be reversed to compensate the apparent energy losses that arise owing to the viscous term. Hence, the reverse trajectory really shows the energy gain during the trajectory reversibility (Figure 5B). However, such an MD simulation becomes extremely unstable (especially for values of γ 0.05) and an explosive behavior of the system was systematically observed after the trajectory was reversed.

Conclusion

Using a simple time reversibility analysis of Hamiltonian systems here we formulated a number of constructive criteria that can potentially be used for evaluating the quality of MD simulations. The analysis of the obtained criteria allowed us, in particular, to select the most appropriate thermostats for a particular molecule under study. For example, in the series of

computational experiments for a polyalanine peptide (FsA peptide), the highest conformational memory was provided by the «no thermostat» scheme. At the same time for most of the studied molecules, the most conformationally stabilizing was the Berendsen's thermostat with the constant τ of 100 fs. Thus, the investigation of satisfiability of the obtained numerical criteria can be a tool both for studying properties, diagnostics and debugging of the MD systems and protocols. On the contrary, systems with the worst degree of conformational memory can be used for more efficient sampling the possible conformations of molecules (Langevin dynamcs). Thus, a new plethora of approaches to making MD thermostat models can be proposed on the basis of the energy conservation law.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Russian Foundation for Basic Research [grant numbers 18-54-00037_Bel_a, 17-04-02105], and Russian Academy of Sciences [Presidium Program in Molecular and Cellular Biology].

ORCID

Ivan Yu. Torshin http://orcid.org/0000-0002-2659-7998

References

Batyanovskii, A. V., & Vlasov, P. K. (2008). Short protein segments with prevalent conformation. *Biophysics*, *53*(4), 264–267. https://doi.org/10.1134/S0006350908040040

Batyanovskii, A. V., Volotovsky, I. D., Namiot, V. A., Filatov, I. V., Galkin, I. A., Gnuchev, N. V., Tumanyan, V. G., & Esipova, N. G. (2015). Structural and physicochemical characteristics of conformationally stable α-helical oligopeptides. *Biophysics*, *60*(3), 348–360. https://doi.org/10.1134/S0006350915030033

Beeman, D. (1976). Some multistep methods for use in molecular dynamics calculations. *Journal of Computational Physics*, *20*(2), 130–139. https://doi.org/10.1016/0021-9991(76)90059-0

Bozdaganyan, M. E., Lokhmatikov, A. V., Voskoboynikova, N., Cherepanov, D. A., Steinhoff, H.-J., Shaitan, K. V., & Mulkidjanian, A. Y. (2019). Proton leakage across lipid bilayers: Oxygen atoms of phospholipid ester linkers align water molecules into transmembrane water wires. *Biochimica et Biophysica Acta. Bioenergetics*, *1860*(6), 439–451. https://doi.org/10.1016/j.bbabio.2019.03.001

Hu, X., Hu, S., Wang, J., Dong, Y., Zhang, L., & Dong, Y. (2018). Steered molecular dynamics for studying ligand unbinding of ecdysone receptor. *Journal of Biomolecular Structure & Dynamics*, *36*(14), 3819–3828. https://doi.org/10.1080/07391102.2017.1401002

Karlov, N. E., & Kirichenko, N. A. (2011). *Oscillations, waves, structures [Kolebaniya, volny, struktury, In Russ]*. Fizmatlit Publ. ISBN 978-5-9221-0205-6.

Komatsu, N., & Abe, T. (2004). Numerical irreversibility in time-reversible molecular dynamics simulation. *Physica D: Nonlinear Phenomena*, *195*(3-4), 391–397. https://doi.org/10.1016/j.physd.2004.05.004

Landau, L. D., & Lifshitz, E. M. (1976). *Course of theoretical physics, vol. 1: Mechanics.* (3rd ed.). Butterworth-Heinemann. ISBN-13: 978-0750628969, ISBN-10: 0750628960

Levesque, D., & Verlet, L. (1993). Molecular dynamics and time reversibility. *Journal of Statistical Physics*, *72*(3-4), 519–537. https://doi.org/10.1007/BF01048022

Lien, Y. H., Ram Mahato, D., Hoppe-Seyler, F., & Fischer, W. B. (2019). Membrane partitioning of peptide aggregates: Coarse-grained molecular dynamics simulations. *Journal Biomolecular Structure and Dynamics*, *27*(2), 1–9. https://doi.org/10.1080/07391102.2019.1581660

Mohseni, A., Molakarimi, M., Taghdir, M., Sajedi, R. H., & Hasannia, S. (2019). Exploring single-domain antibody thermostability by molecular dynamics simulation. *Journal of Biomolecular Structure & Dynamics*, *37*(14), 3686–3696. https://doi.org/10.1080/07391102.2018.1526116

Nishijima, K. (1963). *Fundamental particles (Lecture notes and supplements in physics)*. W.A. Benjamin. ISBN 0805373993

Pant, P., Pathak, A., & Jayaram, B. (2020). Symmetrization of the backbone of nucleic acids: A molecular dynamics study. *Journal of Biomolecular Structure & Dynamics*, *38*(3), 673–681. https://doi.org/10.1080/07391102.2019.1585292

Prakash, A., Kumar, V., Pandey, P., Bharti, D. R., Vishwakarma, P., Singh, R., Hassan, M. I., & Lynn, A. M. (2018). Solvent sensitivity of protein aggregation in Cu, Zn superoxide dismutase: A molecular dynamics simulation study. *Journal of Biomolecular Structure & Dynamics*, *36*(10), 2605–2617. https://doi.org/10.1080/07391102.2017.1364670

Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., & Skiff, W. M. (1992). UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American Chemical Society*, *114*(25), 10024–10035. https://doi.org/10.1021/ja00051a040

Schlick, T. (2002). *Molecular modeling and simulation* (pp. 480). Springer. ISBN 0-387-95404-X.

Tanner, J. E., & Stejskal, E. O. (1968). Restricted self-diffusion of protons in colloidal systems by the pulsed-gradient, spin-echo method. *The* Journal of Chemical Physics, *49*(4), 1768–1770. https://doi.org/10.1063/1.1670306

Torshin, I. Y. (2004). Computed energetics of nucleotides in spatial ribozyme structures: An accurate identification of functional regions from structure. *TheScientificWorldJournal*, *4*, 228–247. https://doi.org/10.1100/tsw.2004.19

Torshin, I. Y. (2006). *Bioinformatics in the post-genomic era: The role of biophysics*. Nova Biomedical Books. ISBN 1-60021-048-1.

Torshin, I. Y., Batyanovskii, A. V., Uroshlev, L. A., Tumanyan, V. G., Volotovskii, I. D., & Esipova, N. G. (2019). Conformational stability/lability of peptide fragments in the sequence context of amino acids. *Biophysics*, *64*(2), 182–271. http://dx.doi.org/10.1134/S0006509190201800

Torshin, I. Y., Esipova, N. G., & Tumanyan, V. G. (2014). Alternatingly twisted β-hairpins and nonglycine residues in the disallowed II′ region of the Ramachandran plot. *Journal of Biomolecular Structure & Dynamics*, *32*(2), 198–208. https://doi.org/10.1080/07391102.2012.759451

Uroshlev, L. A., Torshin, I. Y., Batyanovskii, A. V., Esipova, N. G., & Tumanyan, V. G. (2015). Disallowed Conformations of Polypeptide Chain Exemplified by the β-bend of the β-hairpin in the α-spectrin SH3-domain. *Biophysics*, *60*(1), 1–9. https://doi.org/10.1134/S0006350915010236

Uroshlev, L. A., Torshin, I. Y., Batyanovskii, A. V., Esipova, N. G., & Tumanyan, V. G. (2019). Predetermined conformations in the bends of polypeptide chains: A geometric analysis. *Biophysics*, *64*(2), 195–202. https://doi.org/10.1134/S0006350919020209

Verlet, L. (1967). Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard – Jones molecules. *Physical Review*, *159*(1), 98–103. https://doi.org/10.1103/PhysRev.159.98

Vora, J., Patel, S., Athar, M., Sinha, S., Chhabria, M. T., Jha, P. C., & Shrivastava, N. (2020). Pharmacophore modeling, molecular docking and molecular dynamics simulation for screening and identifying anti-dengue phytocompounds. *Journal of Biomolecular Structure & Dynamics*, *38*(6), 1726–1740. https://doi.org/10.1080/07391102.2019.1615002

Wang, R. R., Liu, W. S., Zhou, L., Ma, Y., & Wang, R. L. (2020). Probing the acting mode and advantages of RMC-4550 as an Src-homology 2 domain-containing protein tyrosine phosphatase (SHP2) inhibitor at

molecular level through molecular docking and molecular dynamics. *Journal of Biomolecular Structure & Dynamics*, *38*(5), 1525–1538. https://doi.org/10.1080/07391102.2019.1613266

Weiner, S. J., Kollman, P. A., Nguyen, D. T., & Case, D. A. (1986). An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry*, *7*(2), 230–241. https://doi.org/10.1002/jcc.540070216 http://dx.doi.org/10.1002/jcc.540070216

Windshügel, B. (2019). Structural insights into ligand-binding pocket formation in Nurr1 by molecular dynamics simulations. *Journal of Biomolecular Structure & Dynamics*, *37*(17), 4651–4657. https://doi.org/10.1080/07391102.2018.1559099

Xiao, J., Melvin, R. L., & Salsbury, F. R. (2019). Jr. Probing light chain mutation effects on thrombin via molecular dynamics simulations and machine learning. *Journal of Biomolecular Structure & Dynamics*, *37*(4), 982–999. https://doi.org/10.1080/07391102.2018.1445032

Zhang, C., Li, Q., Meng, L., & Ren, Y. (2020). Design of novel dopamine D(2) and serotonin 5-HT(2A) receptors dual antagonists toward schizophrenia: An integrated study with QSAR, molecular docking, virtual screening and molecular dynamics simulations. *Journal of Biomolecular Structure and Dynamics*, *38*(3), 860–885. https://doi.org/10.1080/07391102.2019.1590244