Computer Assisted Methods in Engineering and Science, **19**: 257–276, 2012.

From experimental, structural probability distributions to the theoretical causality analysis of molecular changes

Paweł Daniluk[1,2], Maciej Dziubiński[1], Bogdan Lesyng* [1,2]
[1] *Department of Biophysics, Faculty of Physics, University of Warsaw*
Żwirki i Wigury 93, 02-089 Warsaw, Poland
[2] *Bioinformatics Laboratory, Mossakowski Medical Research Centre*
Polish Academy of Sciences
Pawińskiego 5, 02-106 Warsaw, Poland
* *e-mail: lesyng@icm.edu.pl (corresponding author)*

Marta Hallay-Suszek, Franciszek Rakowski, Łukasz Walewski
Interdisciplinary Centre for Mathematical and Computational Modelling
University of Warsaw, Pawińskiego 5A, 02-106 Warsaw, Poland

A brief overview of causality analysis (CA) methods applied to MD simulations data for model biomolecular systems is presented. A CausalMD application for postprocessing of MD data was designed and implemented. MD simulations of two model systems, porphycene (*ab initio* MD) and HIV-1 protease (coarse-grained MD) were carried out and analyzed. Granger's causality methodology based on a Multivariate Autoregressive (MVAR) formalism, followed by the Directed Transfer Function (DTF) analysis was applied. A novel approach based on the descriptors of local structure was also presented and preliminary results were reported. Casuality analyses are required for a better understanding of biomolecular functioning mechanisms. In particular, such analyses can link physics-based structural dynamics with functions inferred from molecular evolution processes. Current limitations and future developments of the presented methodologies are indicated.

Keywords: causality analysis, signal analysis, local descriptors, alignment, MVAR, Directed Transfer Function, molecular dynamics, porphycene, HIV-1 protease, molecular function, molecular evolution.

1. Introduction

Mechanisms of structural changes in biomolecular systems can either be determined from experimental probability distributions or from physics-based multiscale simulations. Experimental distributions are typically either time-averaged or ensemble-averaged quantities with, however, poorly defined physical, ensemble properties. Crystallographic or NMR "structural preferences" are somehow related to effective (coarse-grained) molecular potential energy functions. Experimental preferences can be described in a form of the mentioned above experimental probability distributions, which can be expressed using either Cartesian or cooperative, internal, curvilinear degrees of freedom. Such distributions are often used in designing procedures of effective inter/intramolecular interaction potentials – typically a simple Boltzmann-type relationship is used.

Relationships between experimental distributions ones become much more complex when accounting for time-dependent processes, and when asking what are causal relations between structural transitions. Detecting causal relations in structural changes of biomolecular systems from experimental or theoretical molecular dynamics (MD) simulation data is of crucial importance for the description of molecular mechanisms and understanding the logic of their function. In this study, we concentrate on the analysis of theoretical probability distributions. Time-resolved experimental measurements are typically limited to a few degrees of freedom, whereas theoretical molecular dynamics simulations provide much more detailed knowledge of the biomolecular structures and their dynamics. One may ask the question which kind of data is more reliable. However, since in this study we are focused on methodological aspects of the causality analysis (CA), theoretical simulation data are much more useful. One should also note that current MD simulations are quite accurate, in particular when they are based either on first quantum principle, or on well balanced atomic interaction potentials. Nevertheless, the presented methodology is quite general, and it can be applied to any set of data containing information about time-dependent structural changes. It is based on the signal analysis. There exist some limitations, which will be pointed out, and discussed in more detail. Data obtained from theoretical, time-dependent atomic positions, momenta, forces or their functions can be treated as signals. Quantum degrees of freedom can also be incorporated in the analyses. Assuming that the simulation results precisely predict time-evolution of a given molecular system, we can get to know "how" the system behaves – still we would like to "why" it behaves as "observed". This problem is of particular importance for biomolecular systems determined by molecular evolution mechanisms. These mechanisms tell us "how", and the physics-based description should tell us "why". However, because of complexity of biomolecular or nano systems, determination of causal relations from microscopic interactions is difficult, if not impossible. Therefore, causal relations have to be extracted based on the signal analysis, using methodologies developed in studies of complex systems. An overview of the novel approach to the causality analyses of structural transformations based on MD simulation data and using the signal analyses is given in [20, 21]. In these studies references to signal analyses in biomedical (like EEG) or economic complex systems are also given. CA methodologies, developed and applied in these areas, are similar to some extent.

For the reasons mentioned above, we have been developing a computational platform, CausalMD, for the time-series analysis extracted from MD simulations. Applications of methodologies to selected biomolecular systems are discussed. The first system is porphycene – a molecule in which motions of protons involved in two strong, intramolecular hydrogen bonds are coupled to motions of the molecular scaffold. The second system is HIV-1 protease, the enzyme molecule in which motions of molecular flaps covering the binding site are coupled to motions of other molecular fragments. A brief overview of selected signal analysis methodologies is presented and applied to the mentioned above model systems. Based on the signal analysis, causal relations between molecular motions are described. Applications and future research strategies are indicated.

2. Methodological aspects. Autoregression model, Granger's causality analysis and direct transfer function (DTF)

A theoretical approach for determining an estimator of a directional information flow that is going applied in this study arises from the Granger causality concept. In order to estimate the causal influence of one time series against the other, first of all a model linking and correlating time series under study is introduced. In the simplest one-dimensional case, the signal value in the time t is approximated by the linear combination of p previous values of the same time series:

$$x(t) = \sum_{i=1}^{p} a_i x(t-i) + \epsilon(t), \tag{1}$$

where $\epsilon(t)$ is the noise (presumably white), standing for misfitting of this linear model. This approach to the modeling of the time series is named autoregressive model (AR). In the case of a multidimensional system with several time series describing of its dynamics, it is possible to extend the model into linear combination of the values pertaining to the other signals describing the studied system. This is the so-called multivariate autoregressive model (MVAR). Formally, the model can be presented in the matrix version of the above Eq. (1):

$$\mathbf{X}(t) = \sum_{i=1}^{p} \mathbf{A}_i X(t-i) + \mathbf{E}(t). \tag{2}$$

At this stage, the causal relation are not included. The most important thing is the construction of the multidimensional parameterization of the model, contained in the **A** matrix. It has been proven that such a model reflects well the different features of different data types. As the measure of the model efficiency (namely how well are the data reproduced by the parameter set – entries of the $\mathbf{A}_i$ matrices), a Schwartz-Bayes Criterion has been proposed. This is a heuristic criterion, based on the balance between the reduction of the rest variance (the variance of the noise signal $\epsilon(t)$) and the number of the parameters used in the model:

$$SBC(p) = 2\log(det(\mathbf{V})) + \frac{2kp\log(N)}{N}, \tag{3}$$

where $\mathbf{V} = \langle \mathbf{E}\mathbf{E}^T \rangle$, k is the dimensionality of the system and p is the model order.

The MVAR Eq. (2), can be transformed into the frequency domain, f, and rewritten in the form:

$$\mathbf{A}(f)\mathbf{X}(f) = \mathbf{E}(f), \tag{4}$$

finally using the transfer matrix $\mathbf{H}(f)$ one gets:

$$\mathbf{X}(f) = \mathbf{A}^{-1}(f)\mathbf{E}(f) = \mathbf{H}(f)\mathbf{E}(f). \tag{5}$$

The Granger causality measure, e.g., the causality between channels j (source channel) and i (sink channel) is equivalent to the $H_{ij}(f)$ element of the **H**. The causality strength is a function of frequency f, and the causal function of f is named directed transfer function (DTF). The normalized DTF describes the ratio between the inflow from the channel j to i, to all inflows to channel i:

$$\gamma_{ij}^2(f) = \frac{|H_{ij}(f)|^2}{\sum\limits_{m=1}^{k} |H_{im}(f)|^2}. \tag{6}$$

Such a normalization relates the causal coupling of a chosen pair of channels to the overall information flow (dependence on the selected channel vs. general dependence on the entire set of channels).

Sometimes it is also useful to employ a general measure of causality which is not dependent on frequency:

$$J_{ij} = \int\limits_{f} \gamma_{ij}(f) df. \tag{7}$$

The coefficients J_{ij} and function γ_{ij}^2 will be used as basic causality measures in the following sections.

3. Anharmonic hydrogen motions in porphycene (PC)

3.1. Tautomerization in PC

Hydrogen transfer in molecules is as fundamental to the understanding of chemical and biochemical reactions as is the concept of charge transfer in metallic systems for the design of electronic devices. Many examples of both spontaneous and photo-induced hydrogen transfer reactions can be found in systems ranging from small, isolated molecules [42, 50] to cellular macromolecules such as enzymes or pigments (e.g., bacteriorhodopsin).

Porphycene is a prominent member of this wide class of systems due to an ultra-fast tautomerization reaction that involves a concerted transfer of two hydrogen atoms along the internal N-H$\cdots$N hydrogen bonds (see, e.g., [53] for review). Consequently, the porphycene molecule undergoes continuous inter-conversion between two chemically equivalent, yet distinguishable [52], conformational states roughly billion times per second [18]. The conversion is fully coupled to other intramolecular events, such as vibrational excitations, which enhance or reduce the rate at which the protons move, depending on the excitation energy [51]. It is also sensitive to the interactions with the environment. As demonstrated recently [17] in alkylo-substituted porphycene, the internal conversion can be hindered by steric interactions between substituents and solvent molecules. Vast amount of spectroscopic data was gathered under various experimental conditions, ranging from molecular crystals [30, 55] through organic solvent solutions [37, 54] and isolated molecules [38] down to cryogenic matrices [13, 14, 28, 43], supersonic jet expansions [39] and superfluid helium nanodroplets [49]. Much has been done to interpret the measurments [2, 3, 10, 22, 29, 32, 33, 40, 44]. For example, the long-lasting debate on the reaction mechanism has been concluded with the statement that the concerted transfer of both protons prevails over the step-wise mechanism [18, 29, 41], at least in the room-temperature regime [56, 57].

Relevant to the present study is the rapid development of efficient and accurate computational methods [36] that paved the way for realistic, time-resolved simulations of such molecules as porphycene, with atomistic resolution. In particular, *ab initio* molecular dynamics (AIMD) method is well-suited to studying molecular processes, such as hydrogen transfer, due to its ability to produce real-time trajectories of atomic positions while retaining chemical reactivity as dictated by the underlying electronic structure. Such molecular trajectories provide extensive database of highly correlated data that calls for a reliable analysis tool, which would be capable to extract useful information from the tangle of signals. In the following, we demonstrate how *molecular signals* can be obtained from MD trajectories and how causality analysis can be applied to the resulting time series to assist understanding of anharmonic hydrogen motions in porphycene.

3.2. Extracting signals from AIMD trajectories

In the course of ab initio molecular dynamics (AIMD) simulation Cartesian coordinates of N atoms $\mathbf{R} = \{\mathbf{R}_1, \dots, \mathbf{R}_N\}$ are sampled at constant time intervals Δt from the classical real-time evolution of the molecule. Collection of M such configurations $\mathbf{R}(t),\ t = (\Delta t, 2\Delta t, \dots, M\Delta t)$ forms a molecular trajectory of length $M\Delta t$. Usually, one is interested in certain parameters derived from such a trajectory to describe chemical reactions, conformational changes and other processes. Such parameters are defined as functions of atomic positions $f(\mathbf{R})$, e.g., inter-atomic distances, bond angles, occupation numbers, to name but a few. Standard approach concentrates on time evolution of those parameters alone, or on correlations between few of them (typically two). By contrast, CA as applied here, aims at finding linear correlations between all possible pairs of parameters that describe time evolution of the molecule.

Before applying the CA formalism it is convenient to transform the trajectory from Cartesian to normal mode coordinates, which guarantee that signals have well defined mean value. This is accomplished by projecting the atomic displacements $\Delta\mathbf{R} = \mathbf{R} - \mathbf{R}_0$ from the equilibrium position $\mathbf{R}_0$ onto the normal mode vectors $\mathbf{e}_i,\ i = (1, \dots, 3N)$,

$$p_i(t) = \widehat{\mathbf{M}}\boldsymbol{\Delta}\mathbf{R}(t) \cdot \mathbf{e}_i. \tag{8}$$

Projections $p_i(t)$ are weighted ($\widehat{\mathbf{M}}$ is the $3N \times 3N$ matrix with square roots of atomic masses $\sqrt{m_i}$ on the diagonal), to assure that the kinetic energy is invariant upon transformation. Projected normal modes (PNM) defined in Eq. (8) can be directly used as signals in the causality analysis.

In the present case, the Born-Oppenheimer molecular dynamics simulations of porphycene in the gas phase were carried out using the CPMD code [1]. Ground state electronic structure was treated at the DFT level of theory. Gradient corrected exchange-correlation functional (BLYP) was used [5, 31]. The valence electron wave function was expanded in the plane wave (PW) basis up to 70 Ry, and the interactions of core electrons were described using the norm conserving pseudopotentials [48]. The Poisson equation was solved using the Hockney method [23] in a simple cubic simulation cell with the side length of 15 Å. Decoupling of the periodic images for the electrostatic interactions was applied using cluster boundary conditions [35]. The third order predictor corrector extrapolation of the wave function was applied to decrease the number of SCF cycles required to converge wave function gradients down below 5×10^{-6} a.u. Nuclear equations of motion were integrated using the symplectic velocity Verlet algorithm with a time-step of 20 a.u. (~ 0.48 fs). Center of mass motion was subtracted every 0.5 ps. The total of 45 independent trajectories of 20 ps each was sampled from the canonical ensemble [34] at 300 K. Equilibrium structure (the *trans* form) was optimized and the normal mode analysis was carried out, yielding 108 normal mode vectors.

3.3. Analyzing signals generated by a molecule

The MVAR and DTF procedures were applied using the two-channel model for the following two sets of time series as signals:

$$\begin{aligned} &s^i_{101}(t) := \{p_i(t), p_{101}(t)\},\\ &s^i_{102}(t) := \{p_i(t), p_{102}(t)\}. \end{aligned} \tag{9}$$

$p_i(t)$ is the ith projected normal mode (defined in Eq. (8)), $i = 7114$ enumerates the normal modes of the molecule (108 single vibrations after neglecting the first six normal modes – three translations and three rotations). The modes #101 (Fig. 1a) and #102 (Fig. 1b) are symmetric and asymmetric N-H stretching mode, respectively. We took 20 stationary fragments generated from the 45 AIMD simulations. Based on the time-series analysis we are going to find these degrees of freedom (directions), which influence the broadening of the N-H stretching modes (#101, #102) obtained in experiments. For each pair ($s(t)$) we optimised the order of the model p using the Schwartz Bayesian criterion (SBC). The strongest peak of the distribution of the SBC values for the MVAR analysis was at $p = 11$. As p determines the upper limit of the time correlations in the system, the DTF calculations were carried out for a few different $p's$ in order to check the sensitivity of the results on this parameter. We probed the following values of p: 11, 13, 15 and 31 chosen according to the highest occupied bins. As there was no qualitative change of the results with the increase of p, we selected $p = 11$ as the optimal order of the model in the further part of this work. We computed the normalized DTF ($\gamma(f)$) and the corresponding coefficients J for the two-channel models, $s^i_{101}(t)$ and $s^i_{102}(t)$. Figure 2 shows the influence of the p_j on the p_{101} and the p_{102} ($J_{101,j}$ and $J_{102,j}$, respectively) as the function of the number of the second channel $\#j$. To select the modes that influence selectively the #101 or the #102 mode, the difference between the integrated DTFs, $J_{101,j} - J_{102,j}$ for each $\#j$, in the function of $\#j$ was plotted. Analyzing the plot we can distinguish two groups of modes: the first one that influences the #101 mode significantly stronger than #102, and the second one which influences #102 stronger. The selected modes are listed in the Table 1 with their frequencies and symmetries. The assignments for vibrational modes are taken from [16], where the most of the computed vibrational modes of porphycene have been assigned to the observed ones. In the selected modes we can find four pure in-plane skeletal deformations (12 AG, 8BU, 9 AG, 11 AG), and a number of other types of deformations. As one can see, the symmetry type is conserved. The next step was to recalculate the DTFs for the three-channel model, where the signals were defined as

$$\widetilde{s}^i(t) := \{p_i(t), p_{101}(t), p_{102}(t)\}, \qquad i = 19, 25, 45, 49, 50, 55, 59, 62. \tag{10}$$

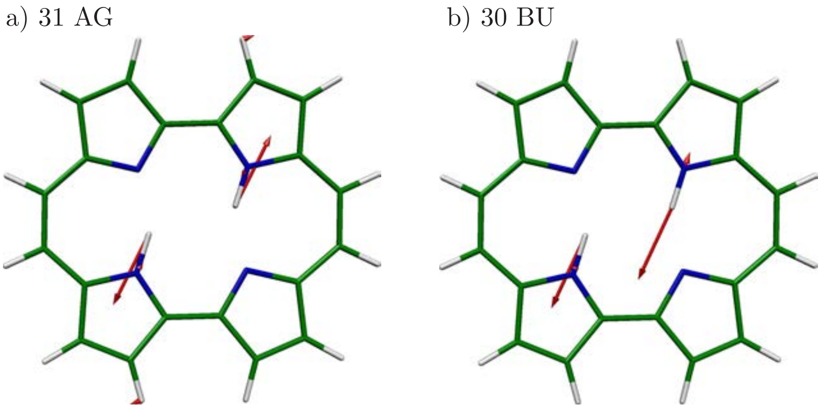

Fig. 1. Displacement vectors of the symmetric (a) and asymmetric (b) N-H stretching vibrations in porphycene that contribute respectively to the concerted and the step-wise mechanisms of the double proton transfer reaction; harmonic frequencies of both normal modes are predicted at the B3LYP/6-31G(d,p) level of theory to yield 2895 cm^{-1}, despite numerous efforts none of them was measured experimentally (see [16] for summary of the experimental issues and recent theoretical analysis).

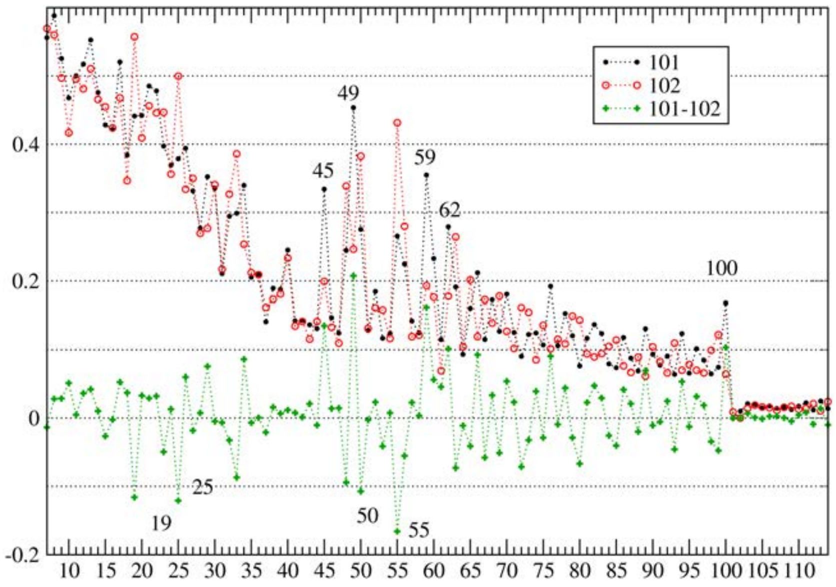

Fig. 2. The comparison of J_{ij} as a function of the number of the first channel index in the two-channel model.

Table 1. The selected vibrational modes, which exhibit significant differences in the influence on the #101 and the #102 modes.

Mode	Freq [cm^{-1}]	Symm	101	102
19	323	2 BU		✓
25	472	4 BU		✓
45	832	13 BG	✓	
49	877	9 AG	✓	
50	896	8 BU		✓
55	947	9 BU		✓
59	994	11 AG	✓	
62	1016	12 AG	✓	
100	1663	30 AG	✓	
101	2895	31 AG		
102	2895	30 BU		

The exemplary DTF function charts are presented in Fig. 3. We chose the two PNMs, which gave the largest discrepancies between the γ values, i.e., p_{49} (a) and p_{55} (b). The first one influences more the #101 mode, and the latter – the #102 one. The value of integrated DTF, J_{ij} (given in the matrix shown in Fig. 4), results from the integration of the function illustrated in the same position as in Fig. 3. The larger value of the integrated DTF, the stronger causal relations between signals. Only two elements from the each chart are significant: γ_{21} and γ_{31}. In Fig. 5, we present the significant elements of γ, that is: γ_{21} and γ_{31} for the six selected PNMs. It can be noticed that the position of the main peak is correlated with the frequency of the first signal in the three-channel model.

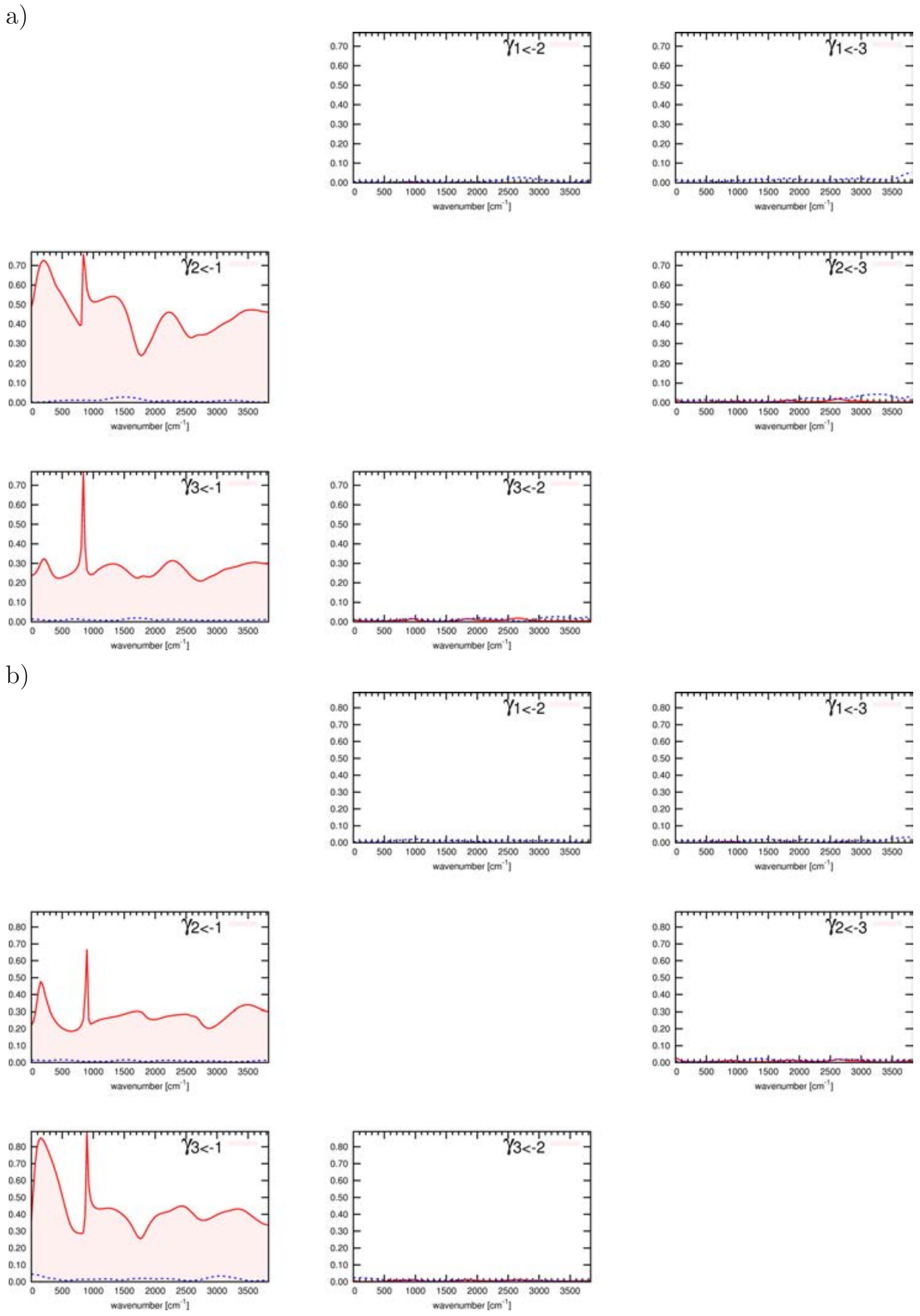

Fig. 3. a) The normalized DTF ($\gamma_{ij}(f)$) for every index pair (i,j) for the three-channel model $\widetilde{s}^{49}$ in the frequency domain, b) the normalized DTF ($\gamma_{ij}(f)$) for every index pair (i,j) for the three-channel model $\widetilde{s}^{55}$ in the frequency domain.

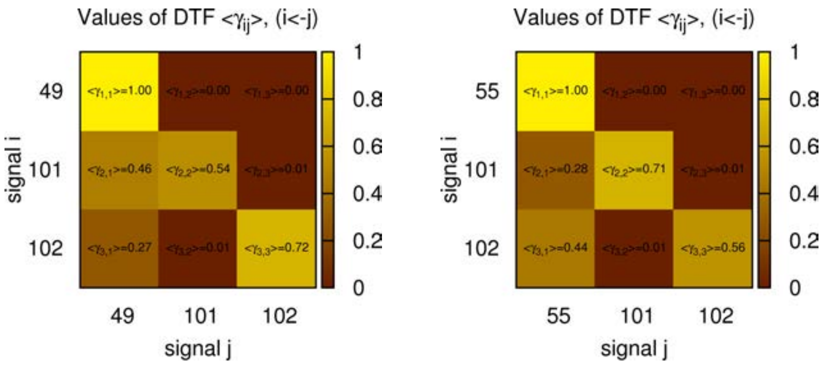

Fig. 4. Results from the integration of the functions DTF, J_{ij} – located in the same positions as in Fig. 3. Note, that the left and right figures correspond to the a) and b) ones in Fig.3, respectively. The larger J_{ij} values in the off-diagonal positions the larger influence of the channel j on the channel i. On the left – the mode 49 influences the modes 101 and 102. On the right – the mode 55 influences the modes 102 and 101.

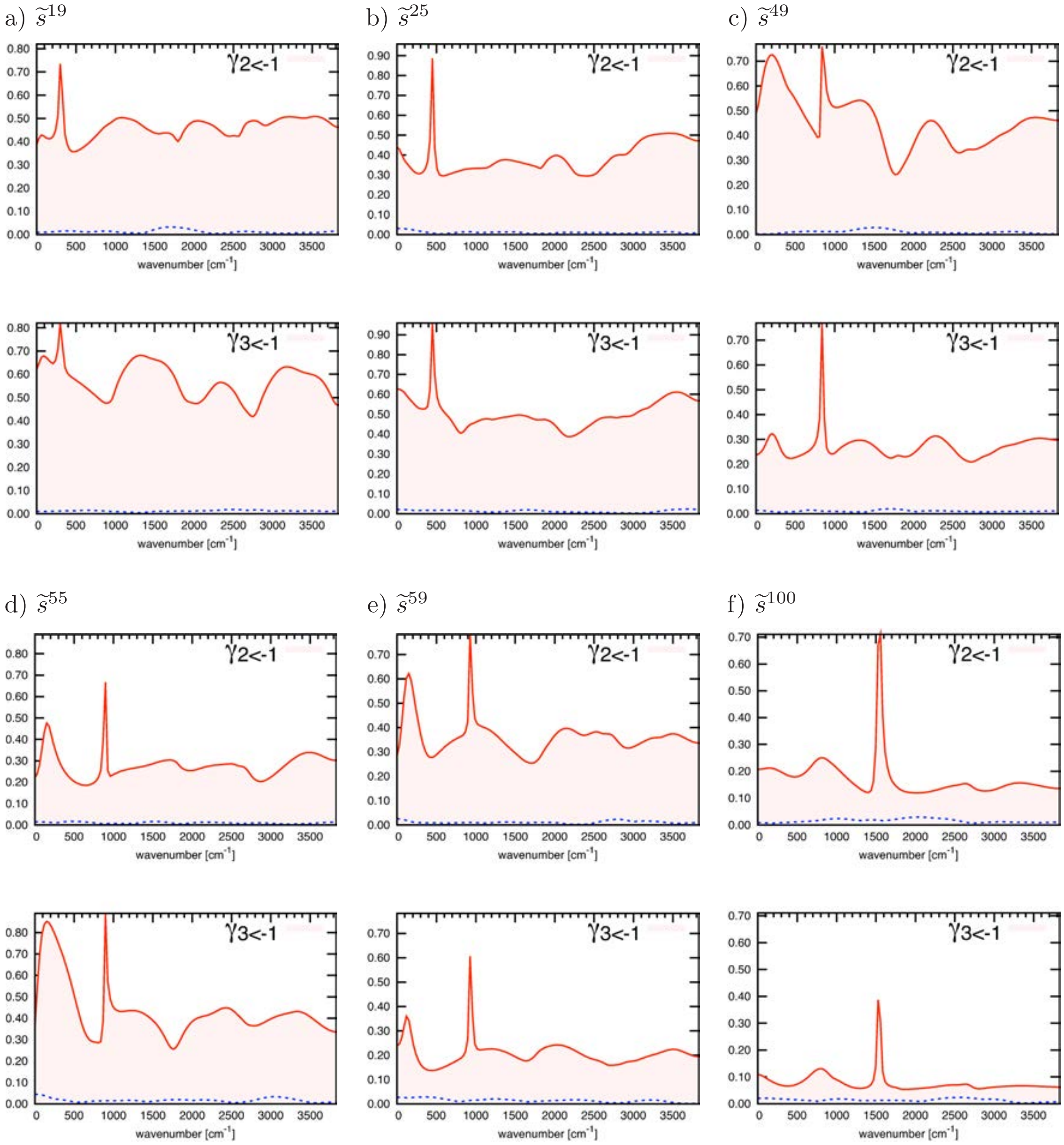

Fig. 5. Selected elements from the γ matrixes, γ_{21} and γ_{31}, computed for the three channel models, $\widetilde{s}^{i}(t)$, where p_i is taken as the first signal and where i denotes the selected PNM.

In conclusion, the analyses presented above were capable to indicate causality relations between a number of modes, indicating most likely some energy flow between them. If the modes are purely harmonic ones, they should behave totally independently. Since they are coupled, a kind of non-linear, physical coupling must exist, and the analysis tells us where such non-linear couplings should be looked for.

4. Causal relation in the protease dynamics

4.1. HIV Protease structure and function

Human immunodeficiency virus type protease (HIV-1 PR) plays an important role in the replication of the virus. Activation of the enzyme proceeds throughout the binding of ligand to the binding side, which is controlled by the motions of the flexible flaps [8]. This mechanism, as a possible target for the anti-virus drugs, was extensively studied in many theoretical and experimental works, see, e.g., [9, 47]. One of the most detailed theoretical descriptions of the HIV-1 protease dynamics, was carried out with the usage of a coarse-grained semi-empirical force field, implemented in the RedMD molecular dynamics package [47]. Simulations carried out in the time-scale of microseconds, in different temperatures, had proven that this simulation method reflects structural changes and thermodynamic properties observed experimentally.

The simulation in the same paradigm was repeated in this study. The simulation was carried out using the Langevin dynamics approach, which is given by the following equations of motion:

$$m_i \frac{d^2 r_i}{dt^2} = F_i(t) - \gamma m_i \frac{dr}{dt} + R_i(t), \tag{11}$$

where m_i is the mass of the residue, F_i is the force, γ is the friction coefficient (collision frequency), and $R_i(t)$ is the random force statistically balancing the friction.

The trajectory, which constitutes the base for the further analysis presented in this study, was obtained with the following parameters: $\gamma = 2$, time step $\triangle t = 20$fs. The geometry was dumped every 1 ps, which gave a large-time resolution in the causal studies. The simulation length was 1 microsecond, in which 17 flap openings were observed. The mean times for the open and closed states were 1.2 ns and 30 ns, respectively.

One of the fundamental features of the HIV-1 dynamics is the presence of the transitions between two thermodynamically quasi-stationary states, which correspond to two free energy minima with flaps open and flaps closed (see Fig. 6 for the definition of degrees of freedom selected for analysis).

Since the mechanism of the flap opening is crucial for the activity of the enzyme, we addressed the question: what is the mechanism, in therms of causality, of such a behavior.

4.2. In search of the causal relations

Causal relations between the elements of dynamical system were systematically examined. Groups of methods which have been used, depend on the concept of the Granger causality. In this approach, causality is defined as a constant, long-winded in time influence of one or more signal channel on other ones. The method does not provide evidence for the causal relations between two, events, one following the other localized in time and separated from each other in the time scale. It tells us rather more about the information flow from one channel to the other. In such a perspective, one may study a causal relation in the dynamical system, as the relation between constantly interacting subsystems or variables in terms of directional flow, e.g., the behavior of one variable or subsystem is partially enslaved by the other.

The dynamics of the HIV-1 protease might exhibit causal properties, but variables taking part in such a causal relation are unknown. The study of such variables constitutes a significant challenge. According to the main goal of this study, which is the research of the flap opening mechanism, the set of the examined variables contains, in each of the analysis, the distance between the flaps, denoted as number 1 in Fig. 6. The search for the interacting degrees of freedom, in the causal sense, was performed in selected fields of mechanical properties of the system. The most obvious category is a set of meaningful distances between residues. Of special interest is the distance between two loops of the hinge region denoted in Fig. 6 by number 2 (and number 3 in the symmetrical chain, because the HIV-1 protein consists of two symmetrical polypeptide chains).

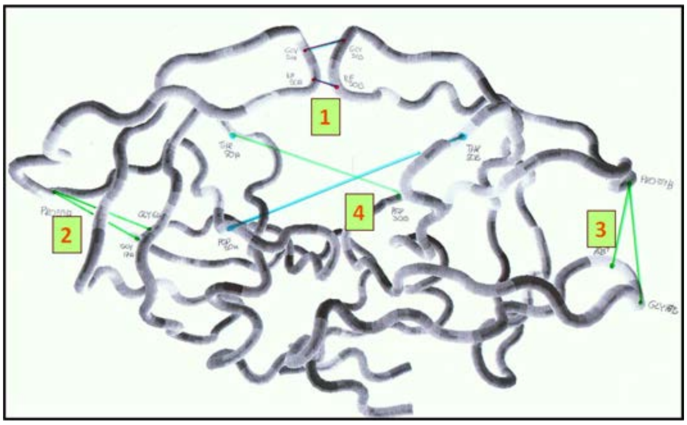

Fig. 6. The HIV protease structure. Distances used in the MVAR modeling are indicated.

Since the mechanical system and its dynamics are driven by forces acting on each residue, we used selected variables related to the forces (mean absolute value, projections on selected “directions”, angular coordinates) as the second category of the studied signal channels. The third category consisted of a set of coefficients, which define the projection of the trajectory of the entire molecule on selected principal directions. Principal directions for the trajectory were computed based on the principal component decomposition (the so-called (PC) essential dynamics) [4]. As it was reported [47] the first and the seventh principal components are involved in the flap opening movement, and the first eigenvector explains substantial amount of the variation of the flap trajectory.

In order to find the causal relation, multivariate autoregressive and directed transfer functions methods were used [7, 27]. The second one gives the insight into the information flow processes in the frequency resolution, namely for each pair of channels and each directed flow a function of frequency is computed. Multivariate Autoregressive model is an application of the Granger causality model to multi-channel systems. As mentioned, the result of such an analysis is the causality coefficient which reflects the strength of a causal relation between channels.

Since the time scale of the causal relations in HIV-1 protease is not known, a various sampling frequencies were tested. The highest sampling frequency was defined by the geometry dumping frequency, performed during the simulations, and is equal to 1 ps. Four next sampling frequencies were used: 5 ps, 10 ps, 20 ps, and 50 ps.

4.3. Causal patterns in the HIV-1 protease

Efficiency of a signal modeling can be tested using the defined above Schwartz-Bayesian criterion (SBC function). The optimal model order is the one where SBC function reaches the minimum, or the beginning of the bottom plateau. Cases which result with non-decaying SBC function shouldn’t be taken into account. Since the autoregressive model is suitable for modeling stationary signals, we can study causal flows in periods of stationarity defined as either open and closed states, separately, or non-specific. The latter one must be long enough to convey several transitions, that might be treated as one stationary signal, in which changes of its value related to the transitions are to be averaged out.

In the first approach presented in this study, the specific signal related to the open state was analyzed. The signal was composed of four channels, namely the flap distance, the left hinge distance, the right hinge distance and the bottom clamp distance. In order to guarantee the proper relation of parameters of the model to the number of data points, several realizations of the same process have been taken into account. Since the signal did not exhibit any spectral properties, it was poorly modeled by the linear autoregressive model. This is also shown in Fig. 7 on the graph presenting SBC criterion function.

In Fig. 7 the result of application of the DTF method to the open conformation periods of the HIV-1 dynamics is shown. Value of the causality flow is not significantly higher than the reference significance level, obtained by hundred-fold bootstrapping of the data.

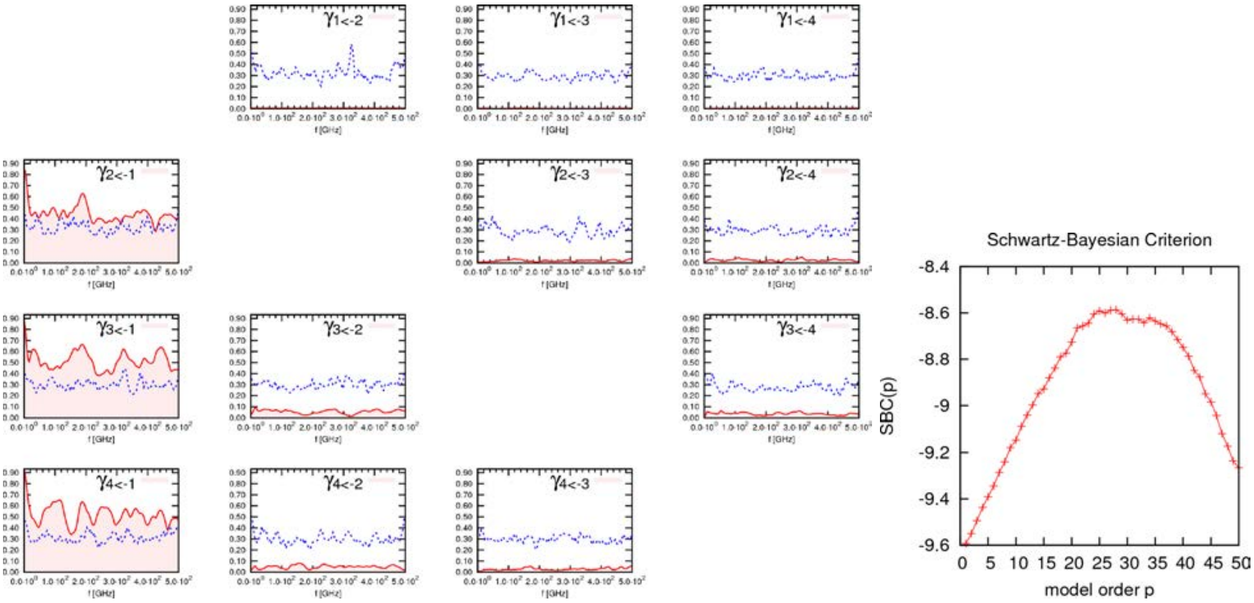

Fig. 7. DTF functions for the four-channel model with variables indicated in Fig. 6. Stationary fragments of the MD trajectories corresponding to the open state of HIV-1 protease were used in the analysis. The blue dashed line denotes the random reference level (p-value = 0.05). The modeling of the signal was not sufficiently successful in this case, as shown on the right graph.

In Fig. 8, the second analysis is presented. The signal is composed of the same channels as previously, but the non-specific time epoch was taken. The first signal (first column in Fig. 8), which corresponds to the distance between flaps, exhibits a causal influence on the rest of the signal taken into the analysis. The strong coupling visible in the low frequency region putatively corresponds to the flap opening process, which persist in the time scale of 40 ns.

In the third approach, see Fig. 9, besides the flap-distance the projection coefficients on the first and the seventh component, and forces acting on the selected residue in the hinge region were additionally taken into account. The information flow pattern shows that first PC (channel #5) constitutes the source of the variability of all others signals. Since the first principal component is related to the flap opening process this is consistent with the commonly known mechanism of HIV-1 protease.

As it was shown, the approach assuming the multivariate autoregressive modeling of the signal, together with DTF has encountered some difficulties when interpreting the results. Nevertheless, the DTF functions presented in Fig. 8 point out that the degree of freedom #1 influences the degrees of freedom #2, #3 and #4. In the third case (Fig. 9), a weak influence of the degree #5 on the degrees #1, #2, #3 and #4 and #6 is visible.

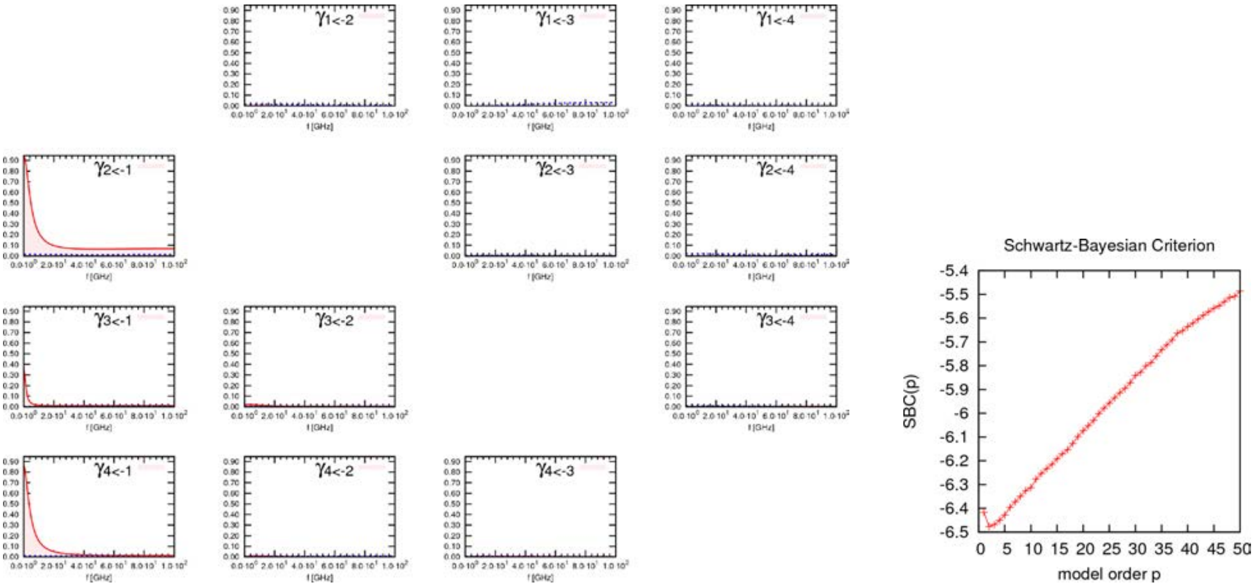

Fig. 8. The DTF functions for the same channels as in Fig. 7. Stationary fragments of the MD trajectories corresponding to both, the open and close states, were analysed. SBC criterion indicates that the model order should equal 2.

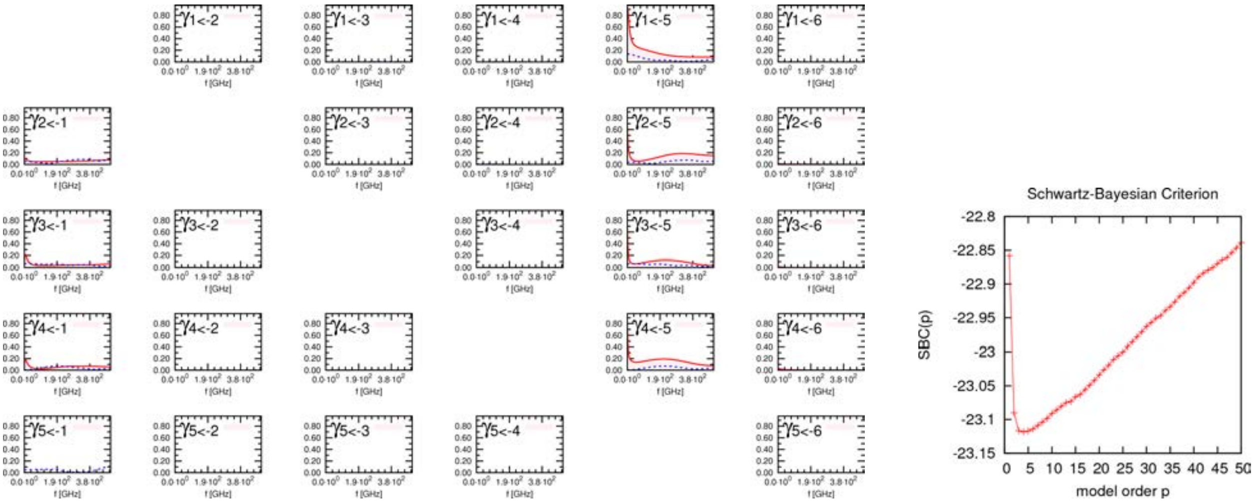

Fig. 9. The DTF functions in the six-channel model. The following channels are taken into account: flap distance (#1), three cylindrical components of a force acting on the flap hinges (#2, #3, #4), as well as the first and the seventh PC coefficients (#5 and #6, respectively). SBC criterion indicates that the model order should equal 4.

5. Descriptors

5.1. Local descriptors of protein structure

Below, a novel prototype approach based on local descriptors – not discussed until now in literature, is presented. A local descriptor is a small part of a structure that can be viewed as a residue-attached local environment. In principle, it is possible to build a descriptor for every residue of a given protein. This process begins by identifying all residues in contact with the central residue descriptors. Elements are then built by including two additional residues along the main-chain, both upstream and downstream of each contact residue. Any overlapping elements are concatenated into single segments. Thus, a descriptor is typically built of several disjoint pieces of the main chain (Fig. 10). It reflects approximately the range of local, most significant physico-chemical interactions between its central residue and other amino-acids. This constitutes a significant difference compared to single segments so frequently used in other studies. Single segments reflect features along the main-chain, while descriptors are spatial, and thus add a three-dimensional context to the local properties of proteins.

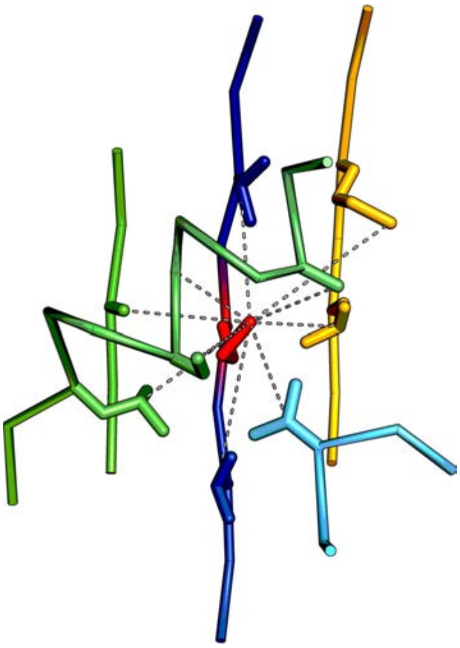

Fig. 10. A sample descriptor (built around the residue `MET70` of the SCOP domain `d1lg7a_`). This descriptor contains nine contacts (dashed lines) between its central amino-acid (red) and residues forming the centers of its elements. Some elements overlap forming longer segments (in particular, fragments of two β-strands (blue and yellow) and fragment of α-helix (green)). Altogether, this descriptor comprises five continuous segments.

Descriptors have already been applied in several studies [6, 12, 15, 24, 25, 45, 46]. Here we use an improved version of the local descriptor methodology described in [12]. Every descriptor is built around its *central amino-acid*. In the first step, we identify residues close to the central amino-acid. For each pair of residues we compute distances between C_α atoms (d_α) and geometrical centers of side-chains R_C (d_C) (for glycine $R_C = C_\alpha$, and for alanine $R_C = C_\beta$.). If either $d_\alpha \leq 6.5$ Å, or $d_C \leq 8$ Å and $d_\alpha - d_C \geq 0.75$ Å (second condition favors residues whose side-chains point towards each other), we consider two residues to be in contact. In the second step, we build *elements* around selected residues by taking four sequential neighbours, two on each side. Finally, overlapping elements are merged into *segments*.

Descriptors may be applied in many aspects of protein structure analysis and prediction, such as prediction of tertiary structure [15] or inter-residue contacts [6] from sequence, prediction of function [24] or ligand binding affinity [45, 46] as well as computing structural alignments [12] and multiple alignments [11].

During MD simulations, a local structure descriptor may, or may not, change its shape. As it will be revealed in further sections, comparing local descriptors that encompass the same residues in different spatial conformations originating from MD trajectory frames is easier. In such a case, the correspondence of residues is trivial and it is sufficient to compute RMSD.

6. Local descriptors of MD trajectories

We used the formalism of local structure descriptors to analyze MD trajectories with respect to two crucial aspects:

1. Identification of states (typically called *the conformational analysis*).
2. Detection of causal relationships between certain regions of the protein (understood as: finding essential structural changes which enable protein transitions between states).

We present the general idea of descriptors of MD trajectories, exemplary data and methods for analyzing such data.

6.1. Definitions

For each of the N configurations generated from an MD simulation, and for each residue, we constructed a descriptor.

We will denote by $A \subset \mathbb{N}$ the set of amino-acids of the whole protein. Let $\mathcal{D}_l^i$ be the set of amino-acids of a descriptor, built around the l-th residue, in the i-th trajectory frame (configuration). During the simulation, the set of residues encompassed by a descriptor varies. The *trajectory descriptor* will be also built around a central; however, we need to take under consideration how often a given residue is a member of the descriptors $\{\mathcal{D}_l^i\}_{i=1}^N$. The **membership function** $\chi_l \colon A \to [0,1]$, for the trajectory descriptor built around the l-th amino-acid, is defined as follows:

$$\chi_l(a) = \frac{\sum\limits_{i=1}^{N} \mathbb{1}_{\mathcal{D}_l^i}(a)}{N}, \tag{12}$$

where $\mathbb{1}_{\mathcal{D}_l^i}$ is a characteristic function of the set $\mathcal{D}_l^i$, defined by

$$\mathbb{1}_{\mathcal{D}_l^i}(a) = \begin{cases} 1, & \text{if } a \in \mathcal{D}_l^i, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

A **trajectory descriptor** $\mathcal{D}_l(f)$ with a membership threshold f is defined as

$$\mathcal{D}_l(f) = \{a \in A \colon \chi_l(a) > f\} \tag{14}$$

i.e. $\mathcal{D}_l(f)$ is a set of residues with memberships higher than the threshold f. Note, that for $f = 0$,

$$\mathcal{D}_l(0) \equiv \bigcup_{i=1}^{N} \mathcal{D}_l^i. \tag{15}$$

6.1.1. Trajectories with different number of steps

Notice that once we choose f, the set $\mathcal{D}_l(f)$ still depends on the length of the trajectory, N. However, for long trajectories ($N > 10000$) this set of amino-acids is (for our purposes) sufficiently invariant.

6.2. Identification of local states

Once we determined $\mathcal{D}_l(f)$, we began the actual analysis of MD trajectory.

A known issue in the field of interpreting molecular simulations, is the problem of identifying the meta-stable states of a given protein. A *state* is, simply put, a subset of microstates of the protein, which are mutually easily accessible. Usually, the definition of "accessibility of microstates" determines the definition of the state. However note, that the definition given here is simplified. A full definition would also need to account for metastability; that's, the requirement that once the protein reaches a given state, the probability that it will stay in that state is sufficiently high.

Difficulty in the identification of states lies in the fact that proteins are very complex systems. A configuration of a n-atom protein requires $3n$ coordinates in the Cartesian coordinate system. If we used RMSD[1] in order to describe the accessibility of microstates we would need to assume that structurally similar configurations are always easily accessible. For high dimensional systems this assumption is not correct.

In our approach, we focus on the *identification of local states*, i.e., for each residue l we identify the states in the system described by $\mathcal{D}_l(f)$. In other words, we divide the protein into intersecting subsystems and analyze their trajectories, in order to identify their (local) states.

The core idea is to define a similarity measure between trajectory descriptors originating from different trajectory frames. Note that descriptors are significantly smaller than the whole protein, and therefore the implication between the structural similarity and a accessibility of microstates is more probable. Therefore, the use of RMSD in the definition of the similarity measure seems justified. Let d_i and d_j be trajectory descriptors obtained from the i-th and j-th trajectory frame. We used a similarity measure p in the following form:

$$p(d_i, d_j) = \exp[-\mathrm{RMSD}(d_i, d_j)/\sigma], \tag{16}$$

where σ is a parameter.

Once we define the similarity measure, we may consider a *similarity matrix* $\mathbf{P}$ defined by $[\mathbf{P}]_{ij} = p(d_i, d_j)$. In graph theory, the similarity matrix is one of the representations of a graph. By analogy, the problem of identifying local states may be looked at as the problem of finding clusters in a graph. Basically, the idea is to find an optimal cut that would group (cluster) vertices of the graph. There are several definitions of an "optimal cut", each of which leads to a different clustering algorithm. We chose a spectral algorithm proposed by Shi and Malik [26].

The spectral algorithm we used in our case is based on the eigenvector analysis of a matrix derived from the similarity matrix. We used hierarchical clustering; therefore, in each step we only needed to consider division into two groups. It is worth noting that spectral clustering involving partitioning into two clusters requires only the second eigenvector, and we used the second eigenvector later in order to determine non-redundant signals.

6.3. HIV-1 protease

We tested our method by analyzing an MD trajectory of the HIV-1 protease. The simulation was carried out using the RedMD force field [19]. In this trajectory, we observed two distinct states of the protease – *flaps opened* and *flaps closed* (Fig. 12).

First, we obtained memberships of residues of protein to the descriptors. We set $f = 0$, and determined trajectory descriptors. Next, for each trajectory descriptor we constructed a similarity matrix with a similarity function defined in Eq. (16).

As a result of the clustering, we obtained vertices of the graph ordered in such a manner, that the first k_1 vertices belong to the first cluster, following k_2 vertices belong to the second cluster, and so on (Fig. 11).

We used clustering the results and labeled each microstate of the trajectory descriptor with the number of the cluster it was attributed to. We then considered the time evolution of the descriptor as a time series of states. Figure 13 presents time series obtained from two trajectory descriptors, one centered on the 24th and the second centered on the 70th amino-acid.

It is worth noting that method we detected two distinct states comprising the "flaps closed" state. Microstates from these two states are apparently mutually accessible, but differ structurally.

We gathered time series for all of the trajectory descriptors. However, most of the time series were redundant and in order to eliminate this redundancy we considered correlations between the

[1] A commonly known measure of divergence, used in the comparison of two protein structures.

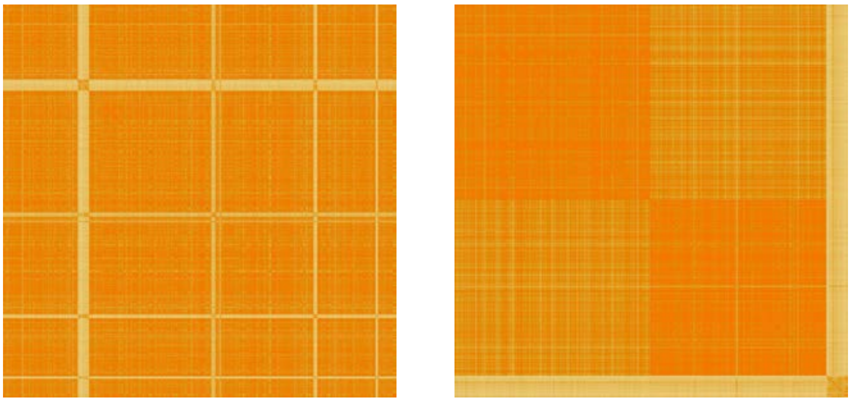

Fig. 11. On the left – a similarity matrix of a selected trajectory descriptor obtained from the HIV-1 protease MD simulation. Orange indicates high and yellow – low similarity. The matrix on the right presents the result of the clustering where vertices have been reordered and we see that three clusters are found in the graph.

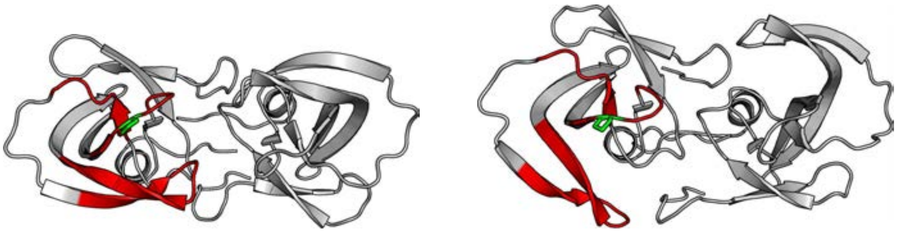

Fig. 12. Two trajectory frames of the whole protein, HIV-1 protease. On the left is an example of a configuration attributed to the "flaps closed", and on the right – to the "flaps opened" state.

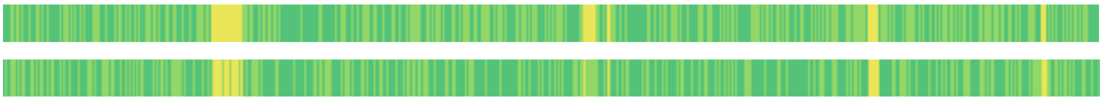

Fig. 13. Graphical representations of time series (i.e., sequences of states acquired from analyses) of two trajectory descriptors. In both cases three states were identified, hence in both depictions we used three colors (green, pale-green and yellow).

second eigenvectors acquired from the spectral clustering. Figure 14 presents correlations between the vectors from the analysis of the 195 trajectory descriptors. We used an absolute value of the Pearson correlation.

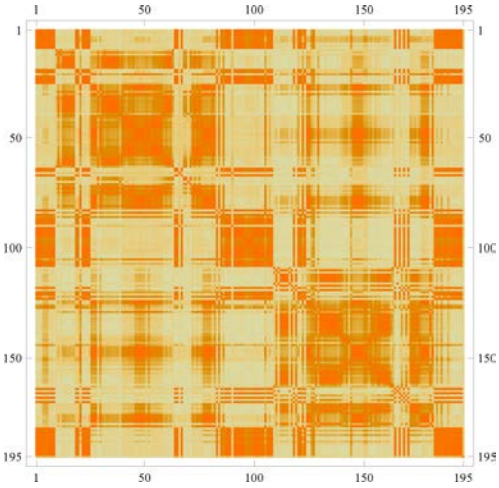

Fig. 14. Correlations between the second eigenvectors.

6.4. A new thread for the causality analysis

A natural question arises: “Is there a sequence of microstates which leads to the *flaps opened* state?”, or in other words: “Is there a causal relation between the sequence of microstates and the event of opening the flaps?”

Our first approach to this problem was an attempt to determine if there are any shifts occurring in the time series of states. To discover such shifts we carried out alignments of the time series using a commonly known Needleman-Wunsch algorithm. Figure 15 presents an alignment of the previously shown time series. Our concept was that precedence in the time series might imply causality between the subsystems described by the corresponding trajectory descriptors. Figure 16 presents shifts in the time series resulting from the alignment presented previously.

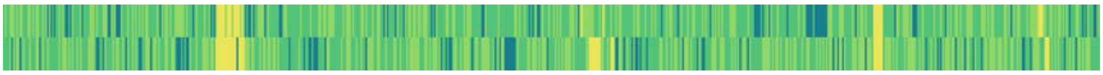

Fig. 15. Alignment of the time series presented in Fig. 13. The dark-green-colored gaps were inserted to allow for optimal matches of the corresponding clustered states.

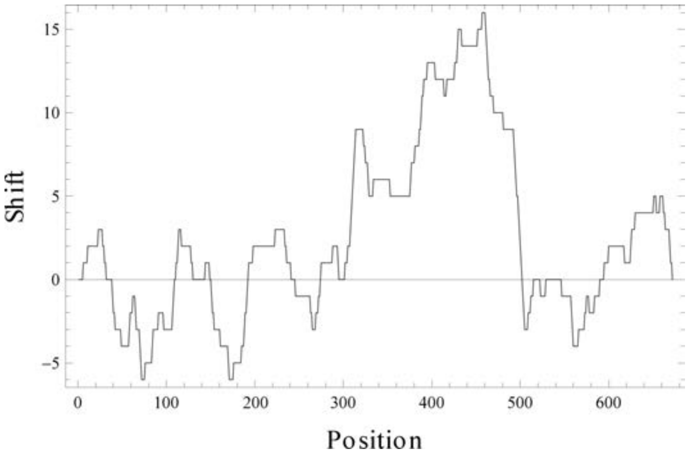

Fig. 16. Shifts in time series alignment.

This shows that trajectory descriptors are ideal candidates for finding causal relations within the protein. We imagine that in order to carry out its function, proteins need to undergo conformational transitions. Such transitions are enabled by “hooks” and “hinges”, which, only for a set of combinations, lead to a successful transition between states.

For example, the HIV-1 protease in order to conduct proteolysis needs to be in the “flaps opened” state. The active site is then accessible. If we could find a sequence of events leading to the transition from the “flaps closed” to the “flaps opened” state, we could locate regions whose activity is crucial in this process. The detected regions can also be prone to inhibition. This may open a new prospect for a more advanced pharmaceutical research.

6.5. Future work

The presented results should be considered to be a preliminary proof of concept. One of the issues that remain to be solved is identification of corresponding local states. In case of larger and more complex system, it may happen that the number of identified local states will vary between descriptors. Also, there might not exist a strict correspondence (such as $A_1 \leftrightarrow B_2$, $B_1 \leftrightarrow A_2$, $C_1 \leftrightarrow C_2$). Nevertheless, it might be possible to infer less detailed relationships (e.g., that A_1 and B_1 exclude C_2). Such relationships might be inferred by finding a scoring matrix which maximizes the score of the alignment between respective time series.

Autoregressive model although applicable to many kinds of time series with values in real numbers, should not be applied to discrete data, such as states of a descriptor. Therefore, it seems necessary to use a nonlinear model in this case. We intend to design an artificial neural network, and use it in the Granger formalism.

7. Conclusions

Understanding of functioning mechanisms of complex systems, in particular, biomolecular or nano-systems requires applications of formal causality analyses. An overview of causality analysis (CA) methods applied to MD simulations data of selected biomolecular systems was presented. A CausalMD application was designed and implemented. In this analysis, time-dependent positions, distances, forces, momenta, projected normal modes (PNM) or their functions are treated as signals. MD simulations of two model systems, porphycene (ab initio MD) and HIV-1 protease (reduced MD) were carried out and analysed. The Granger causality methodology based on the multivariate autoregressive (MVAR) formalism followed by the directed transfer function (DTF) analysis were applied. Causality relation coefficients referring to correlated motions of selected degrees of freedom were computed and analyzed. It was pointed out, that some degrees of freedom in the systems under the study are in causal relations. This provides valuable information about functional properties of these systems. In the case when PNM amplitudes are treated as signals, the presented analysis indicated that some normal modes are in causal relations. If these were pure harmonic normal modes their time-evolutions should be independent – this results from basic physical principles. The existing correlations simply show that these are not pure harmonic modes, and nonlinear couplings exist between them. The presented post-processing indicates possible physical couplings, which can be identified in details with further physics-based analyses. Finally, a new approach, based on the local structure descriptors was presented and preliminary results were reported. This approach can be applied to larger biomolecular systems – it requires, however, further studies, in particular, finding optimal criteria for the alignment of time-dependent sequences of the descriptors. This can be done by designing of so-called "substitution matrices", known from conventional sequence analysis. However, this is not a simple problem, and it requires further, advanced studies. Extending the present MVAR/DTF model from scalar signals to vector ones is another direction of future research. Preliminary developments has already been carried out in our group using a quaternion formalism. One should also point out that the presented methodology applies precisely only to stationary signals with constant mean values. In practical applications, MD signals, particularly in shorter time-frames, are not strictly stationary, which requires further methodological improvements.

8. Acknowledgements

Authors thank Dr. Joanna Trylska for her help in carrying out MD simulations for the HIV-1 protease using RedMD, as well as for valuable discussions, and thank Piotr Majka for his contribution in coding the CausalMD application. These studies were supported by the Ministry of Science and Higher Education grant (N N519 384736). MD simulations were carried out in the NIC Supercomputing Centre, Juelich, and at ICM, University of Warsaw. Analysis of the MD simulations data was done using the POIG.02.01.00-14-122/09 computing infrastructure at Faculty of Physics, University of Warsaw, as well as using the infrastructure of Biocentrum-Ochota (POIG.02.03.00-00-0030/09).

References

[1] CPMD. http://www.cpmd.org/. Copyright IBM Corp 1990-2008, Copyright MPI für Festkörperforschung Stuttgart 1997-2001.

[2] Geometric H/D Isotope Effects and Cooperativity of the Hydrogen Bonds in Porphycene. *Chem. Phys. Chem.*, **8**: 315–321, 2007.

[3] M. Abdel-Latif, O. Kühn. Laser control of double proton transfer in porphycenes: towards an ultrafast switch for photonic molecular wires. *Theor. Chem. Acc.*, **128**: 307–316, 2011. DOI: 10.1007/s00214-010-0847-y.

[4] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen. Essential dynamics of proteins. *Proteins: Structure, Function and Genetics*, **17**(4): 412–425, 1993.

[5] A.D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, **38**: 3098–3100, Sep 1988.

[6] P. Björkholm, P. Daniluk, A. Kryshtafovych, K. Fidelis, R. Andersson, T.R. Hvidsten. Using multi-data hidden Markov models trained on local neighborhoods of protein structure to predict residue-residue contacts. *Bioinformatics*, **25**(10): 1264–70, 2009.

[7] K.J. Blinowska, R. Kuś, M. Kamiński. Granger causality and information flow in multivariate processes. *Phys. Rev. E*, **70**: 050902, Nov 2004.

[8] C. Chang, T. Shen, J. Trylska, V. Tozzini, J.A. McCammon. Gated binding of ligands to HIV-1 protease: Brownian dynamics simulations in a coarse-grained model. *Biophysical journal*, **90**(11): 3880–3885, 2006.

[9] J.R. Collins, S.K. Burt, J.W. Erickson. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. *Nature structural biology*, **2**(4): 334–338, 1995.

[10] H. Cybulski, M. Pecul, T. Helgaker, M. Jaszuński. Theoretical Studies of Nuclear Magnetic Resonance Parameters for the Proton-Exchange Pathways in Porphyrin and Porphycene. *J. Phys. Chem. A*, **109**: 4162–4171, 2005.

[11] P. Daniluk. *Analysis of structural similarity of proteins using local structure descriptors* [in Polish: *Analiza podobieństwa struktur przestrzennych białek przy użyciu deskryptorów lokalnej struktury*]. Ph.D. Thesis, University of Warsaw, 2011.

[12] P. Daniluk, B. Lesyng. A novel method to compare protein structures using local descriptors. *BMC bioinformatics*, **12**(1): 344, 2011.

[13] J. Dobkowski, V. Galievsky, M. Gil, J. Waluk. Time-Resolved Fluorescence Studies of Porphycene Isolated in Low-Temperature Gas Matrices. *Chem. Phys. Lett.*, **394**: 410–414, 2004.

[14] J. Dobkowski, V. Galievsky, A. Starukhin, J. Waluk. Relaxation in Excited States of Porphycene in Low-Temperature Argon and Nitrogen Matrices. *Chem. Phys. Lett.*, **318**: 79–84, 2000.

[15] M. Drabikowski, S. Nowakowski, J. Tiuryn. Library of local descriptors models the core of proteins accurately. *Proteins*, **69**(3): 499–510, 2007.

[16] S. Gawinkowski, Ł. Walewski, A. Vdovin, A. Slenczka, S. Rols, M.R. Johnson, B. Lesyng, J. Waluk. Vibrations and hydrogen bonding in porphycene. *Phys. Chem. Chem. Phys.*, **14**: 5489–5503, 2012.

[17] M. Gil, J. Dobkowski, G. Wiosna-Sałyga, N. Urbańska, P. Fita, C. Radzewicz, M. Pietraszkiewicz, P. Borowicz, D. Marks, M. Glasbeek, J. Waluk. Unusual, Solvent Viscosity-Controlled Tautomerism and Photophysics: Meso-Alkylated Porphycenes. *J. Am. Chem. Soc.*, **132**(38): 13472–13485, 2010.

[18] M. Gil, J. Waluk. Vibrational Gating of Double Hydrogen Tunneling in Porphycene. *J. Am. Chem. Soc.*, **129**: 1335–1341, 2007.

[19] A. Górecki, M. Szypowski, M. Długosz, J. Trylska. RedMD – reduced molecular dynamics package. *J. Comput. Chem.*, **30**(14): 2364–73, 2009.

[20] A. Gorecki, J. Trylska, B. Lesyng. Causal relations in molecular dynamics from the multi-variate autoregressive model. *EPL (Europhysics Letters)*, **75**: 503, 2006.

[21] A. Gorecki, J. Trylska, B. Lesyng. Causality and correlation analyses of molecular dynamics simulation data. In *From Computational Biophysics to Systems Biology (CBSB07)*, volume NIC Series 36, pages 25–30, Juelich, 2007. John von Neumann Institute for Computing.

[22] J. Hasegawa, K. Takata, T. Miyahara, S. Neya, M.J. Frisch, H. Nakatsuji. Excited States of Porphyrin Isomers and Porphycene Derivatives: A SAC-CI Study. *J. Phys. Chem. A*, **109**: 3187–3200, 2005.

[23] R.W. Hockney. The potential calculation and some applications. *Methods Comput. Phys.*, **9**: 135–211, 1970.

[24] T.R. Hvidsten, A. Kryshtafovych, K. Fidelis. Local descriptors of protein structure: a systematic analysis of the sequence-structure relationship in proteins using short- and long-range interactions. *Proteins*, **75**(4): 870–84, 2009.

[25] T.R. Hvidsten, A. Kryshtafovych, J. Komorowski, K. Fidelis. A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics*, **19** Suppl. 2: ii81–91, 2003.

[26] J.M.J. Shi. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**(8): 888–905, 2000.

[27] M. Kaminski, K. Blinowska. A new method of the description of the information flow in the brain structures. *Biological Cybernetics*, **65**: 203–210, 1991. DOI: 10.1007/BF00198091.

[28] C.W.M. Kay, G. Elger, K. Möbius. The Photoexcited Triplet State of Free-Base Porphycene: A Time-Resolved EPR and Electron Spin Echo Investigation. *Phys. Chem. Chem. Phys.*, **1**: 3999–4002, 1999.

[29] P.M. Kozlowski, M.Z. Zgierski, J. Baker. The Inner-Hydrogen Migration and Ground-State Structure of Porphycene. *J. Chem. Phys.*, **109**: 5905–5913, 1998.

[30] U. Langer, C. Hoelger, B. Wehrle, L. Latanowicz, E. Vogel, H.-H. Limbach. ^{15}N NMR Study of Proton Localization and Proton Transfer Thermodynamics and Kinetics in Polycrystalline Porphycene. *J. Phys. Org. Chem.*, **13**: 23–34, 2000.

[31] C. Lee, W. Yang, R.G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, **37**: 785–789, Jan 1988.

[32] K. Malsch, G. Hohlneicher. The Force Field of Porphycene: A Theoretical and Experimental Approach. *J. Phys. Chem. A*, **101**: 8409–8416, 1997.

[33] K. Malsch, M. Roeb, V. Karuth, G. Hohlneicher. The Importance of Electron Correlation for the Ground State Structure of Porphycene and Tetraoxaporphyrin-Dication. *Chem. Phys.*, **227**: 331–348, 1998.

[34] G.J. Martyna, M.L. Klein, M. Tuckerman. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.*, **97**(4): 2635–2643, 1992.

[35] G.J. Martyna, M.E. Tuckerman. A reciprocal space based method for treating long range interactions in ab initio and force-field-based calculations in clusters. *J. Chem. Phys.*, **110**(6): 2810–2821, 1999.

[36] D. Marx, J. Hutter. *Ab Initio Molecular Dynamics*. Cambridge University Press, 2009.

[37] M. Pietrzak, M. Shibl, M. Broring, O. Kuhn, H.-H. Limbach. ^{1}H/^{2}H NMR Studies of Geometric H/D Isotope Effects on the Coupled Hydrogen Bonds in Porphycene Derivatives. *J. Am. Chem. Soc.*, **129**: 296–304, 2007.

[38] H. Piwoński, C. Stupperich, A. Hartschuh, J. Sepioł, A. Meixner, J. Waluk. Imaging of Tautomerism in a Single Molecule. *J. Am. Chem. Soc.*, **127**: 5302–5303, 2005.

[39] J. Sepioł, Y. Stepanenko, A. Vdovin, A. Mordziński, E. Vogel, J. Waluk. Proton Tunneling in Porphycene Seeded in a Supersonic Jet. *Chem. Phys. Lett.*, **296**: 549–556, 1998.

[40] M. F. Shibl, M. Tachikawa, O. Kühn. The Geometric (H/D) Isotope Effect in Porphycene: Grid-Based Born–Oppenheimer Vibrational Wavefunctions *vs.* Multi-Component Molecular Orbital Theory. *Phys. Chem. Chem. Phys.*, **7**: 1368–1373, 2005.

[41] Z. Smedarchina, M.F. Shibl, O. Kühn, A. Fernández-Ramos. The Tautomerization Dynamics of Porphycene and Its Isotopomers - Concerted Versus Stepwise Mechanisms. *Chem. Phys. Lett.*, **436**: 314–321, 2007.

[42] Z. Smith, E. Wilson, R. W. Duerst. The infrared spectrum of gaseous malonaldehyde (3-hydroxy-2-propenal). *Spectrochimica Acta Part A: Molecular Spectroscopy*, **39**(12): 1117–1129, 1983.

[43] A. Starukhin, E. Vogel, J. Waluk. Electronic Spectra of Porphycenes in Rare Gas and Nitrogen Matrices. *J. Phys. Chem. A*, **102**: 9999–10006, 1998.

[44] E. Steiner, P.W. Fowler. The Four-Electron Diamagnetic Ring Current of Porphycene. *Org. Biomol. Chem.*, **1**: 1785–1789, 2003.

[45] H. Strömbergsson, P. Daniluk, A. Kryshtafovych, K. Fidelis, J.E. Wikberg, G.J. Kleywegt, T.R. Hvidsten. Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. *J. Chem. Inf. Model*, 2008.

[46] H. Strömbergsson, A. Kryshtafovych, P. Prusis, K. Fidelis, J.E. Wikberg, J. Komorowski, T.R. Hvidsten. Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. *Proteins*, **65**(3): 568–79, 2006.

[47] V. Tozzini, J. Trylska, C. en Chang, J.A. McCammon. Flap opening dynamics in HIV-1 protease explored with a coarse-grained model. *Journal of Structural Biology*, **157**(3): 606–615, 2007.

[48] N. Troullier, J.L. Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, **43**: 1993–2006, Jan 1991.

[49] A. Vdovin, J. Waluk, B. Dick, A. Slenczka. Mode-Selective Promotion and Isotope Effects of Concerted Double-Hydrogen Tunneling in Porphycene Embedded in Superfluid Helium Nanodroplets. *ChemPhysChem*, **10**(5): 761–765, 2009.

[50] J.W.F. Rowe, R.W. Duerst, E.B. Wilson. The Intramolecular Hydrogen Bond in Malonaldehyde. *J. Am. Chem. Soc.*, **98**: 4021–4023, 1976.

[51] Ł. Walewski, J. Waluk, B. Lesyng. Car-Parrinello Molecular Dynamics Study of the Intramolecular Vibrational Mode-Sensitive Double Proton-Transfer Mechanisms in Porphycene. *J. Phys. Chem. A*, **114**: 10753–10757, 2010.

[52] J. Waluk. Ground- and Excited-State Tautomerism in Porphycenes. *Acc. Chem. Res.*, **39**: 945–952, 2006.

[53] J. Waluk. Tautomerization in Porphycenes. In *Hydrogen Transfer Reactions*, J.T. Hynes, J.P. Klinman, H.H. Limbach, R.L. Schowen [Eds.]. Wiley-VCH, Weinheim, 2007.

[54] J. Waluk, M. Müller, P. Swiderek, M. Köcher, E. Vogel, G. Hohlneicher, J. Michl. Electronic States of Porphycenes. *J. Am. Chem. Soc.*, **113**: 5511–5527, 1991.

[55] B. Wehrle, H.H. Limbach, M. Köcher, O. Ermer, E. Vogel. ^{15}N-CPMAS-NMR Study of the Problem of NH Tautomerism in Crystalline Porphine and Porphycene. *Angew. Chem. Int. Ed. Engl.*, **26**: 934–936, 1987.

[56] T. Yoshikawa, S. Sugawara, T. Takayanagi, M. Shiga, M. Tachikawa. Theoretical study on the mechanism of double proton transfer in porphycene by path-integral molecular dynamics simulations. *Chem. Phys. Lett.*, **496**(1–3): 14–19, 2010.

[57] T. Yoshikawa, S. Sugawara, T. Takayanagi, M. Shiga, M. Tachikawa. Quantum tautomerization in porphycene and its isotopomers: Path-integral molecular dynamics simulations. *Chemical Physics*, **394**(1): 46–51, 2012.