John von Neumann Institute for Computing

Causality and Correlation Analyses of Molecular Dynamics Simulation Data

A. Gorecki, J. Trylska, B. Lesyng

published in

From Computational Biophysics to Systems Biology (CBSB07),
Proceedings of the NIC Workshop 2007,
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **36**, ISBN 978-3-9810843-2-0, pp. 25-30, 2007.

http://www.fz-juelich.de/nic-series/volume36

Causality and Correlation Analyses of Molecular Dynamics Simulation Data

Adam Gorecki[1,2], **Joanna Trylska**[2], **and Bogdan Lesyng**[1]

[1] Department of Biophysics and CoE BioExploratorium, Faculty of Physics, University of Warsaw
Zwirki i Wigury 93, 02-089 Warsaw, Poland

[2] Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw
Pawinskiego 5A, 02-106 Warsaw, Poland
E-mail: {agorecki,joanna,lesyng}@icm.edu.pl

The subject of this study is the application and tuning of existing statistical analysis methods for molecular dynamics (MD) data analysis. Special attention is focused on detecting causality relationships (time precedence) between events in MD, based on time series derived from trajectories. The problems of time-series preprocessing, such as normalization and filtering, and the choise of the most appropriate causality detection method is discussed. Features and characteristics of two existing and widely applied methods: Directed Transfer Function[3] and conventional Granger causality approaches[2] are described. We suggest an adaptation of the conventional Granger method for MD analysis. The adapted Granger method is tested using the MD/SCC-DFTB simulation data of the proton transfer reaction in malonaldehyde[4] and a coarse-grained MD simulation of HIV-1 protease[5].

1 Introduction

Detecting causality relationships between conformational changes in biomolecular systems simulated with molecular dynamics (MD) methods is of crucial importance for describing their mechanisms and understanding the logic of their functioning. An attempt to approach this problem was presented in our recent study[1]. We followed the Granger causality methodology[2] and applied a Multi-Variate Autoregressive Model (MVAR) with Directed Transfer Function(DTF), which was used successfully in EEG time-series analyses[3]. However, the method still requires some tuning, and in this presentation we deal mostly with a conventional Granger approach[2]. We analyse also two following problems - normalization of the data and the noise filtering.

2 The Causality Analysis Model

Classical correlation analysis detects linear coupling between variables at the same time but it cannot detect linear couplings with a time shift or nonlinear couplings. One of the more advanced solutions is the MVAR model, which can detect time-shifted linear couplings:

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

where: $\mathbf{X}(t) = \{X_1(t), \ldots, X_k(t)\}$ - vector of analysed k variables at time t, called also channels; $t - i \equiv t - i \cdot dt$ - a notation for the time shift of i steps backward;

$\mathbf{A}(i),\ i = 1, \ldots, p$ - fitted MVAR coefficients, matrices of $k \times k$ dimension; p - model order; $\mathbf{E}(t)$ - white noise vector of k dimension.

The $\mathbf{A}(i)$ are fitted to satisfy condition (1) and keep $\mathbf{E}(t)$ components, $E_i(t)$, linearly uncorrelated, with proper mean and standard deviation. The standard deviations of $E_i(t)$ correspond to the white noise levels in each variable and determine $\mathbf{A}(i)$ estimation.

The raw MVAR coefficients are usually not representative for our purposes. Results of the MVAR fit usually depend on the model order (real signals do not satisfy a strict MVAR model), they can be also different for subsets of time series. In our purposes the MVAR model plays a role of a searching engine, not the system parametrization method.

Some time ago we tested[1] a more comfortable representation of the MVAR analysis namely, the Directed Transfer Method[3]. This method was designed for EEG analysis and it is based on the frequency representation of signal transmission, well optimized for linear systems with a clear linear filter interpretation, and noise level independent. However, the method requires variables of the same units and a similar signal type (such as electric potentials recorded as EEG signals). It is sensitive to scaling variables and gives ambiguous results for variables expressed in different units. This problem is very important in MD data analysis because MD simulation observables, such as distances, angles, combinations of different degrees of freedom, energies, etc., have different units. Normalization of variables is connected with choosing the appropriate noise level for the rescaled variables, which determines the MVAR fit.

In this study we test an older, but equivalent to DTF, model: the conventional Granger causality[2] approach. This method is based on comparing of the MVAR fit error *with* and *without* selected variable information. To estimate causal influence of X_j variable on variable X_i, we should select MVAR-model order p, and perform the following MVAR fits.

From the fit for full variables set $\mathbf{X}(t) = \{X_1(t), \ldots, X_k(t)\}$

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

we compute the residual variance matrix $\mathbf{V} = \langle \mathbf{E}^T(t)\mathbf{E}(t) \rangle$.
From the fit for variables set with X_j excluded:

$$\mathbf{X}^{(j)}(t) = \{X_1(t), ..., X_{j-1}(t), X_{j+1}(t), ..., X_k(t)\}$$

$$\mathbf{X}^{(j)}(t) = \sum_{i=1}^{p} \mathbf{A}^{(j)}(i)\mathbf{X}^{(j)}(t-i) + \mathbf{E}^{(j)}(t), \tag{3}$$

we compute the residual variance matrix: $\mathbf{V}^{(j)} = \langle (\mathbf{E}^{(j)}(t))^T \mathbf{E}^{(j)}(t) \rangle$.
Note that $\dim \mathbf{V} = k \times k$ and $\dim \mathbf{V}^{(j)} = k - 1 \times k - 1$.
The causality measure for $X_j \to X_i$ direction is defined as:

$$J_{ij} = 1 - \frac{V_{ii}}{V_{ii}^{(j)}} \in [0; 1]. \tag{4}$$

The J_{ij} matrix is usually asymmetric and represents the strength of a delayed linear coupling between pairs of variables $X_j \to X_i$, where 0 corresponds to no coupling, 1 - to X_i fully determined by X_j with the linear relationship.

The conventional Granger method results are not sensitive to rescaling and unit choice, which results from the construction of this method. This feature is very important for our applications. Optional renormalization can be applied for better numerical MVAR fitting convergence and accuracy.

3 The Conventional Granger Method - Examples of Applications

3.1 Malonaldehyde Molecular Dynamics Trajectory

Malonaldehyde molecule is shown in Fig.1a. The analysed trajectory was derived from a combined quantum/classical dynamics of the proton transfer between O1 and O2 oxygen atoms[4], and contains 1000000 observations probed every 10 fs (every 10 steps of dynamics). Previous analyses show that the proton transfer is a cooperative reaction and the

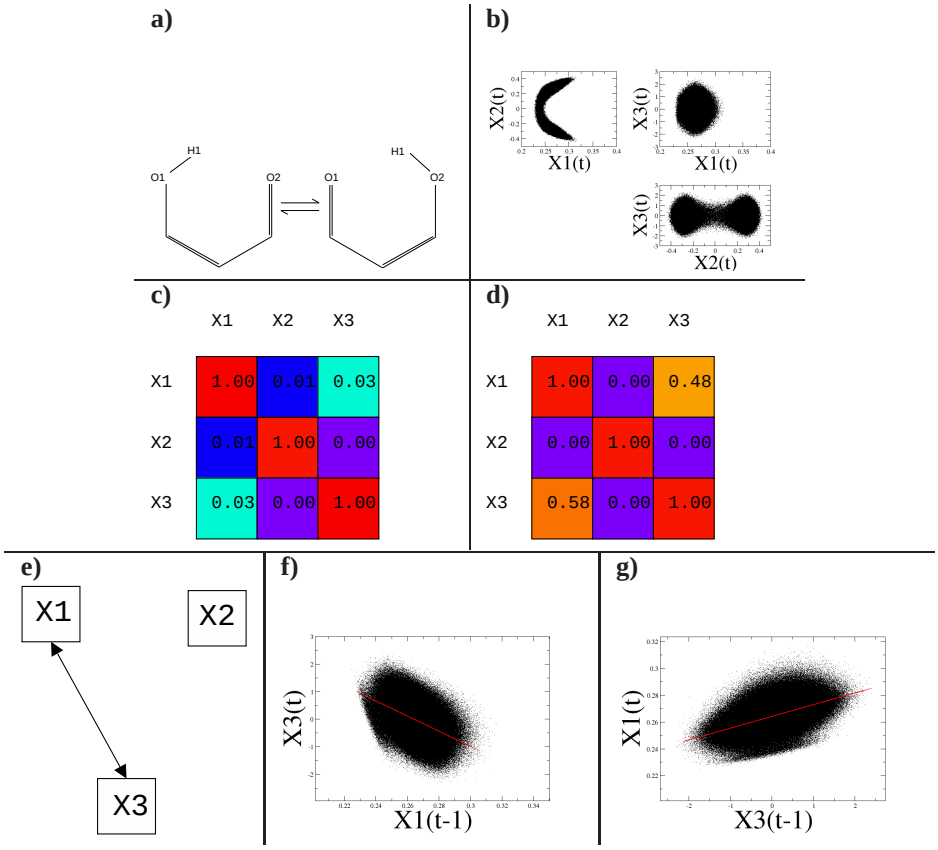

Figure 1. Malonaldehyde example: a) the malonaldehyde molecule; b) relationships between values of X_1, X_2, X_3 variables in the same time t; c) classical correlation matrix for the X_1, X_2, X_3 variables defined in Eqn.5 ; d) the Granger causality matrix J_{ij} for the X_1, X_2, X_3; e) causality diagram corresponding to J_{ij} with the shown feedback $X_1 \leftrightarrow X_3$; f,g) relationships between values of the X_1, X_3 variables with the time shift ± 1.

required condition for the proton hopping is the small O1 and O2 distance. We selected following variables for the causality search:

$$X_1 = |O_1O_2|,\ X_2 = \frac{|HO_1| - |HO_2|}{|HO_1| + |HO_2|},\ X_3 = (v_{\vec{O2}} - v_{\vec{O1}}) \cdot \vec{O_1O_2}/|O_1O_2| \quad (5)$$

The X_2 variable is called the „reaction coordinate" and in our example describes the relative proton position. The X_3 variable is the projection of relative velocities of O1 and O2 atoms on the $\vec{O_1O_2}$ direction. Simple correlation matrix (Fig.1c) does not show any interesting correlation because the variables measured in the same moment are either independent or couplings are nonlinear (Fig. 1b). The Granger method applied to malonaldehyde data shows strong bidirectional causality relationship (feedback) between variables X_1 and X_3. It is an example of existing of non-instant correlations between the spatial and velocity degrees of freedom.For automatic detection of the X_1 and X_2 couplings, we need to apply nonlinear and/or instant causality extensions of the conventional Granger approach.

3.2 HIV-1 Protease Molecular Dynamics Trajectory

We have analysed the coarse-grained dynamics trajectory of HIV-1 protease (Fig.2a) performed in the NVE ensemble[5]. The trajectory contained 10000 frames, delayed by 10 ps. The conformation of the protein was represented by reduced variables: 10 PCA projections derived from the Essential Dynamics method[7]. No instant linear correlations between the variables were seen in this parametrization because they are already included in the PCA eigenvectors. The Granger method detected only weak couplings (Fig.2b) and for the analysis we have selected those with $J_{ij} > 0.1$ (diagram in Fig.2c). J_{16} is the largest element in the first row, which indicates the $X_6 \rightarrow X_1$ coupling. This coupling is strongly nonlinear (Fig.2d) and difficult to detect by linear methods, but the Granger method detected it as a result of some linear correlation. The coupled PCA motions are shown in Fig.2ef - the eigenvector corresponding to X_1 is the most significant movement which characterizes the flap opening. The second one, corresponding to X_6, is also a component of flap movement. Both components change distance between the 17 and 39 residues of two symmetric chains. The movement in X_6 direction is preceding the X_1 flap opening movements, but the Granger method detects it as not very important statistically. A similar correlation was described in the cited articles[5,6].

4 The Influence of Smoothing

Smoothing algorithms (e.g. Savitzky-Golay filter[9]), are usually based on linear filters. The output signal at time t is a linear combination of input values over a window of times: $X_i'(t) = \sum_{k=-k_0}^{k=+k_0} c(k) X_i(t+k)$. Linear filtering of the signal obviously interferes with the MVAR model; the MVAR model will detect filter parameters when run on a too densely probed data. In some cases, filtering can remove some noise from the signal, and can help in detecting of the expected couplings. Sampling of the data for the MVAR analysis should have lower density than the window size used for the preceding smoothing operation.

We applied smoothing of the data by Savitzky-Golay polynomial filter[9], with $k_0 = 4$ (then window was 9 frames long) and order of polynomial 2, then probed data every 10 steps. This operation doesn't improve causal relations visibility.

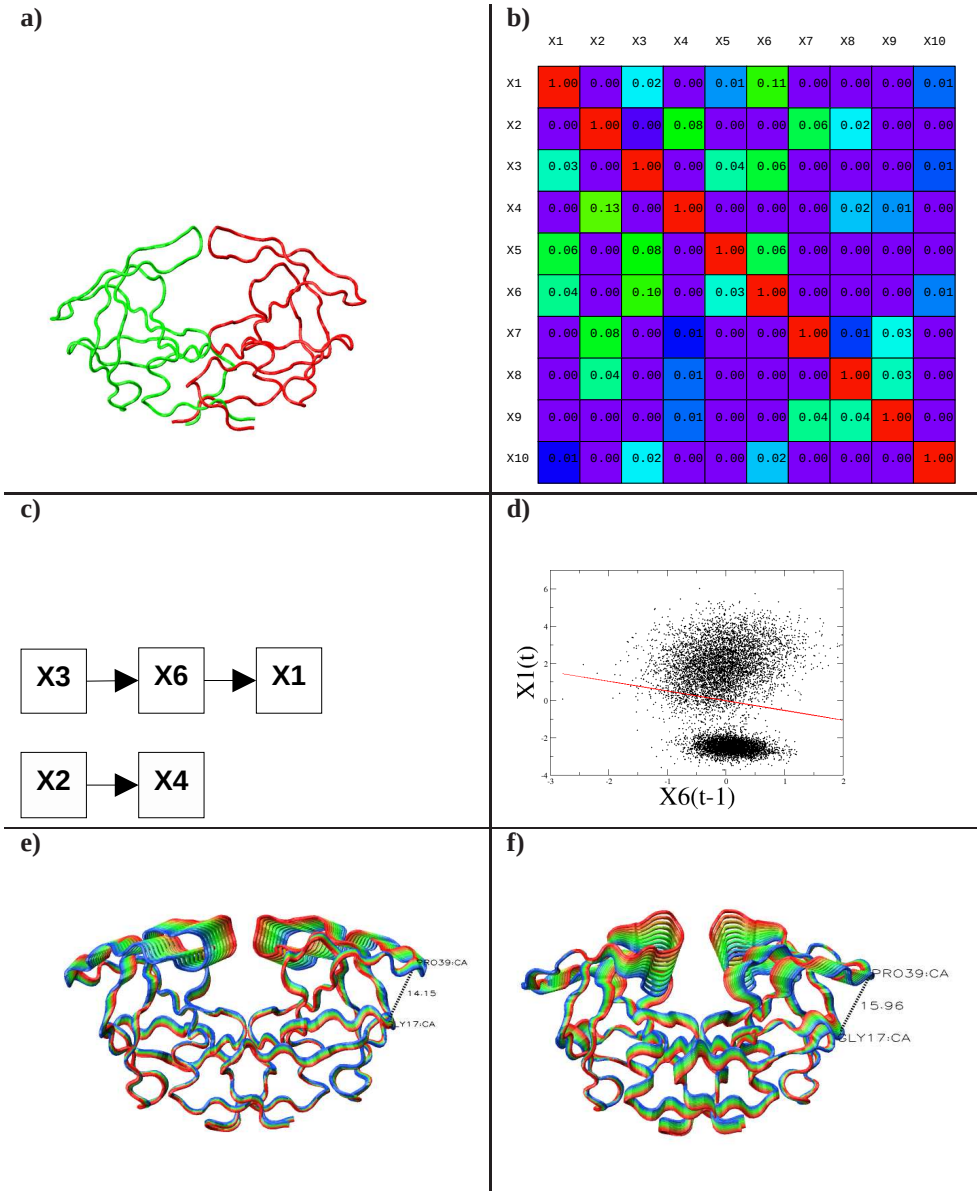

Figure 2. HIV-1 protease example: a) the ribbon model of HIV-1 PR homodimer; b) the Granger causality matrix J_{ij} for the $X_1, \ldots, X_{10}$ variables (reduced coordinates which result from the projection of the MD coordinates on ten most significant PCA components); c) causality diagram corresponding to J_{ij} which shows causality relation for $J_{ij} \geq 0.1$; d) the relationship $X_6 \rightarrow X_1$ with the time shift of 1 e,f) the PCA eigenvectors (movement directions) corresponding to the X_1 and X_6 projections, respectively. Motions occur from blue to red, and back.

5 Conclusions

We have been developing and applying two versions of the causality analyses: the MVAR/DTF and the conventional Granger approaches. MVAR/DTF was presented in[1]. This study deals mostly with the conventional Granger approach. We analysed MD simulation data of malonaldehyde and HIV-1 protease.

The conventional Granger method based on the Multi-Variate Autoregression Model is a quite efficient tool for the molecular dynamic data analysis, because it is independent on normalization and can be applied for signal channels characterized with different units. We have been developing the generalization of this method for detecting of non-linear couplings. For the studied examples prefiltering of the data before the MVAR analysis did not improve the sensitivity of the method in detecting the expected couplings. It interferes with the MVAR and should be use with some care.

Acknowledgments

We would like to acknowledge R. KUS from the Department of Biomedical Physics for providing his very fast MVAR calculation software, L. WALEWSKI of the Department of Biophysics of Warsaw University for providing the molecular dynamics trajectories of proton transfer in the malonaldehyde molecule, V.TOZZINI for providing the molecular dynamics trajectories of HIV-1 protease. We are also grateful to M.KAMINSKI, R.KUS and K.BLINOWSKA from the Department of Biomedical Physics of Warsaw University for helpful discussions and consultations on the applications of the MVAR and DTF methods. The authors were supported by Warsaw University (115/E-343/ICM/BST-1076/2005). JT acknowledges support from University of Warsaw (115/30/E-343/S/2007/ICM BST 1255), Polish Ministry of Science and Higher Education (3 T11F 005 30, 2006-2008), Fogarty International Center (NIH Research Grant # R03 TW07318) and Foundation for Polish Science. AG thanks Polish Ministry of Science and Higher Education (N202 079 32/1841,2007-2008).

References

1. A. Gorecki, J. Trylska, B. Lesyng, Europhys. Let., 2006, 75, 503-509.
2. C. W. J. Granger, Econometrica, 1969, 37, 424-438.
3. K.J. Blinowska, R. Kus, M.J. Kaminski, Phys. Rev. E, 2004, 70, 050902.
4. L. Walewski, P. Bala, M. Elstner, Th. Frauenheim, B. Lesyng, Chem. Phys. Let., 2004, 397, 45 1-458.
5. V. Tozzini, J. Trylska, Chia-en Chang, J. A. McCammon, J. Struct. Biol., 2007, 157, 606-615.
6. V. Tozzini, J. A. McCammon, Chem. Phys. Lett., 2005, 413:123-128.
7. A.Amadei, A.B.M.Linssen, H.J.C.Berendsen, Proteins:Struct.Funct.Genet., 1993, 17, 412.
8. R. Hegger, H. Kantz, and T. Schreiber, CHAOS 9,1999 , 413 .
9. Numerical Recipes, chapter 14.8, `http://www.nr.com`.