Biophysical Chemistry 196 (2015) 1–9

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: http://www.elsevier.com/locate/biophyschem

Random Matrix Theory in molecular dynamics analysis

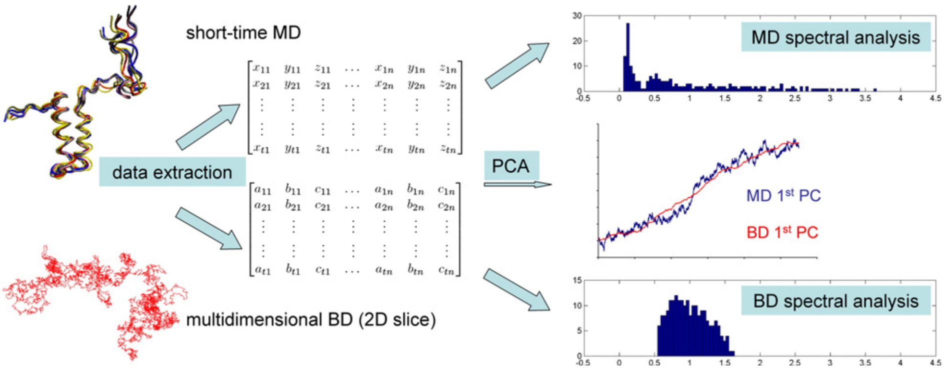
Luigi Leonardo Palese

SMBNOS, University of Bari "Aldo Moro", Piazza G. Cesare, Policlinico, 70124 Bari, Italy

HIGHLIGHTS

- PCA performed on short-time MD experiments leads to cosine-shaped projections.
- Also PCA performed on multidimensional Brownian dynamics leads to the same result.
- We use Random Matrix Theory tools in order to compare MD data with Brownian systems.
- We show that protein dynamics is not really Brownian also at very short time-scale.
- We suggest that Random Matrix Theory can be very useful in MD data analysis.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 6 March 2014
Received in revised form 26 August 2014
Accepted 27 August 2014
Available online 4 September 2014

Keywords:
Molecular dynamics
Principal component analysis
Random Matrix Theory
ApoCox17
Convergence
Brownian dynamics

ABSTRACT

It is well known that, in some situations, principal component analysis (PCA) carried out on molecular dynamics data results in the appearance of cosine-shaped low index projections. Because this is reminiscent of the results obtained by performing PCA on a multidimensional Brownian dynamics, it has been suggested that short-time protein dynamics is essentially nothing more than a noisy signal. Here we use Random Matrix Theory to analyze a series of short-time molecular dynamics experiments which are specifically designed to be simulations with high cosine content. We use as a model system the protein apoCox17, a mitochondrial copper chaperone. Spectral analysis on correlation matrices allows to easily differentiate random correlations, simply deriving from the finite length of the process, from non-random signals reflecting the intrinsic system properties. Our results clearly show that protein dynamics is not really Brownian also in presence of the cosine-shaped low index projections on principal axes.

1. Introduction

Protein functions, such as substrate recognition and release, enzymatic activity and allosteric regulation, require conformational transitions. Due to inherent difficulties to experimentally access to the time-resolved protein motions, molecular dynamics has been increasingly used in the study of molecular conformations in functionally-relevant motions at atomic detail [1]. Nowadays, molecular dynamics protocols permit to obtain accurate prediction of experimental observables (see, for example Ref. [2]). However, the enormous intrinsic dimensionality of biological systems poses serious intelligibility problems. To overcome these difficulties, a series of techniques have been used in order to obtain low-dimensional and meaningful representations of the system dynamics [3,4]. The search for collective coordinate systems, permitting to identify subspaces in which functionally significant protein motions could be easily and accurately identified, is nowadays an active and attractive research field [3]. Among the computational methods for

E-mail address: luigileonardo.palese@uniba.it.

http://dx.doi.org/10.1016/j.bpc.2014.08.007
0301-4622/© 2014 Elsevier B.V. All rights reserved.

identifying useful collective coordinates one of the most widely used is the normal mode analysis [5,6], which is based on the harmonic approximation of the conformational energy surface. However, this approach relies essentially on a single conformation, which is assumed to correspond to the minimum energy structure. The presence of multiple minima in the protein conformational energy landscape has determined a wide use of computational approaches more suitable to be applied on the large number of molecular configurations obtained by molecular (or also Monte Carlo) dynamics. Principal component analysis (PCA) is one of the most popular computational tool used for this task [4,7,8], based both on the mass-weighted covariance matrix, as in the quasi-harmonic analysis [9], or on a non-mass-weighted covariance matrix, which is the approach used in the essential dynamics version [10]. Even if methods able to detect nonlinear correlations in molecular dynamics analysis have been proposed, such as the nonlinear principal component analysis [11,12], the full correlation analysis [13] and the Isomap-based routines [14,15], still the most widely used methods for dimensionality reduction are PCA-based algorithms.

For systems in which a single potential well is an appropriate representation of the conformational energy landscape, the mass-weighted principal modes correspond essentially to the normal modes, i.e. the eigenvectors of the mass-weighted Hessian matrix matching the energy minimum configuration. However, for systems in which multiple minima exist (or are at least supposed), analysis of the non-mass-weighted covariance matrix is more appropriate. By this way, PCA may suitably account for anharmonic molecular motions, thus providing access to the largest collective atomic fluctuations. Generally, more than 80% of the total atomic fluctuations are contained in less than 20% of the principal axes.

One major drawback of covariance matrix-based analyses has been pointed out: these methods critically depend on sampling. As was shown in Ref. [16] principal components from the short-time multidimensional protein simulations are cosine (or sine) shaped, similar to what is observed in multidimensional random diffusion [16,17]. The problem of how to separate intrinsic properties of the molecular system from sampling artifacts has led to a series of studies and proposals [17–20]. Generally, the accuracy of the covariance matrix analysis is considered to depend on the statistical relevance of configuration space sampled within the simulation time-course. Whereby, a number of suggestions have been made in order to evaluate the so called ‘convergence’ of simulations. The cosine content method [16,17,21] or the overlap measures of the essential subspaces [17,19,20], based on the root mean square inner product of the essential eigenvectors, are among the most popular ones. The question of the essential eigenvector convergence has been addressed in several studies [3,18–22].

Here we show that the cosine-shaped appearance of the principal component projections in molecular dynamics analysis does not mean that protein motions are featureless, or equivalent to random diffusion. The physical reason of these cosine-shaped low index projections is simply related to the fact that, in short time-scale, proteins explore a flat landscape, with shallow minima. Here we use a method able to discriminate true non-random dynamics from pure random motions, which is based on the Random Matrix Theory (RMT) [23,24]. This method is suitable for the intrinsic system properties' extraction, also in presence of apparently barrier-less dynamics, such as, even if not limited to, short time-scale simulations.

2. Methods

2.1. Molecular dynamics simulations set-up

Fully reduced apo-Cox17, PDB [25] entry 1U97 [26], has been used as model system, similarly to what was reported in Ref. [27]. The protein was immersed in a water sphere containing 6080 TIP3P type water molecules and five counterbalancing potassium ions to preserve electroneutrality. Molecular dynamics simulations were performed by NAMD [28,29] using the all-atom Charmm22 force field [30] with CMAP correction [31]. Simulations were run at 310 K in the NVT ensemble essentially as described [27]. Each simulation run lasted for 1.1 ns after the minimization and equilibration protocol. Data extraction was done using VMD [32].

For each simulation run $T+1$ conformations were sampled (including the starting one). The extracted data are in the form of atomic position vectors: each vector in the conformational vector set has dimension $N=3n$ and is of the form $x_1, y_1, z_1, \ldots, x_n, y_n, z_n$, where each x_i, y_i, z_i corresponds to the Cartesian coordinates of the ith α-carbon atom. The sampled conformations were arranged in an empirical data matrix of dimension $(T+1)\times N$.

2.2. Principal component analysis and Random Matrix Theory

For data of dimensionality N, PCA permits to compute N so-called principal components (PCs), which are N-dimensional vectors that are aligned with the maximum variance directions of the data. The PCs must form an orthonormal basis, i.e. they are all mutually perpendicular and have unit length, so they are uncorrelated.

In the classical PCA algorithm, the input data consist of $T+1$ observations x_t, each of dimension N. From these observations, a centered matrix is constructed by subtracting the mean value of each degree of freedom time series. By this way we obtain a matrix whose elements are atomic displacements from an average conformation (note that this last conformation does not have a physical significance). The transpose of the displacement matrix can be used for the Pearson's coefficient matrix calculation (see below for details). We use the rank-ordered eigenvectors of the Pearson's coefficient matrix as PCs, instead of the correlation matrix eigenvectors, and projections of the original centered data on the PCs can be done simply by performing the dot product, as usual.

The transpose of the temporal evolution matrix representing the protein α-carbon atoms (see above) can be used to build a position difference matrix D of dimension $N\times T$, whose elements are

$$D_{\alpha t}=x_{\alpha(t+1)}-x_{\alpha t}. \tag{1}$$

From this difference matrix a new matrix X is constructed, whose elements are

$$X_{\alpha i}=\frac{1}{\sigma_\alpha}\left(x_\alpha(i)-\bar{x}_\alpha\right) \tag{2}$$

where σ_α represents the standard deviation of each degree of freedom time series. Based on this matrix, a correlation matrix of size $N\times N$ can be obtained:

$$C=\frac{1}{T}XX^T \tag{3}$$

where the ^{T} means the transpose matrix, and whose elements are the Pearson's coefficients $C_{\alpha\beta}$. Statistical dependencies among the signals (representing the degree of freedom time series) are revealed by the non-zero elements of C. Eigenvalues and eigenvectors of C can be obtained by solving the equation

$$Cv_k=\lambda_k v_k \tag{4}$$

and the usual convention $\lambda_1\geq\lambda_2\geq\lambda_3\geq\ldots\geq\lambda_N$ is applied. Because, by construction, the correlation matrix C is real and symmetric, its eigenvalues λ_k and the corresponding eigenvectors v_k are also real. Note that, since $C_{\alpha\alpha}=1$ we have:

$$\sum_{k=1}^{N}\lambda_k=\mathrm{Trace}(C)=\sum_{\alpha=1}^{N}C_{\alpha\alpha}=N. \tag{5}$$