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

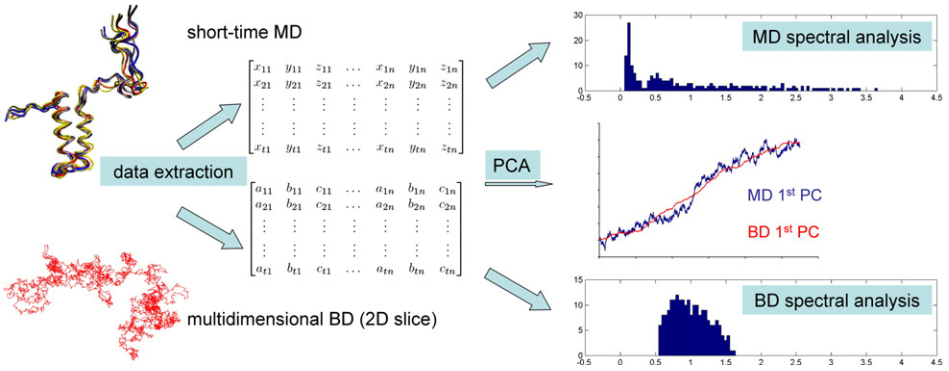
Luigi Leonardo Palese

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/

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

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

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:

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

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

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

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:

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

The matrix V containing the ordered eigenvectors can be used to project the original empirical data on the new basis system simply by computing the dot product $Y = V^T X$.

When looking at the eigenspectrum of a correlation matrix it is essential to distinguish between correlations arising from true dependencies among the system degrees of freedom and false correlations related to the finite length of signals. This task can be done by the RMT [24,33], which offers universal predictions for matrix ensembles with specific properties. Starting from a Gaussian random matrix G_β, i.e. a matrix of independent and identically distributed (i.i.d.) standard random normal elements, a series of random matrix ensembles can be obtained [24]. The G_β entries can be real ($\beta = 1$), complex ($\beta = 2$) or quaternions ($\beta = 4$). If G_β is a $p \times p$ matrix we can describe the Gaussian ensembles as $(G_\beta + G_\beta')/2$. G_β' denotes the transpose of a real matrix ($\beta = 1$), the Hermitian transpose of a complex matrix ($\beta = 2$) or the dual transpose of a quaternion matrix ($\beta = 4$). By this way, we can define the Gaussian orthogonal ensemble (GOE, $\beta = 1$), the Gaussian unitary ensemble (GUE, $\beta = 2$) and the Gaussian symplectic ensemble (GSE, $\beta = 4$). The Gaussian ensembles are of paramount importance in many fields of physics. If G_β is a $q \times p$ ($q \geq p$) matrix, we can define, in general, the Wishart ensembles as $G_\beta' G_\beta$ [24,34]. As for the Gaussian ensembles, the Wishart ones can be represented by $p \times p$ symmetric ($\beta = 1$), Hermitian ($\beta = 2$) or self dual ($\beta = 4$) matrices. The Wishart ensembles, and the related Jacobi/MANOVA ensembles [24], arise naturally in multivariate statistics. Other random matrix ensembles can be defined (the Fourier/circular ensemble). In the case of molecular dynamics analysis, we must consider real and symmetric correlation matrices, so the pertinent ensemble is the Wishart ($\beta = 1$) one. We define our Wishart-type matrix as:

$$W = \frac{1}{T} M M^T \tag{6}$$

where the matrix M contains elements from a zero-centered Gaussian distribution. In our case we construct such random matrices in two different ways. We obtain a Wishart-type ensemble by randomly shuffling, along columns, the empirical simulation matrix whose entries are the α-carbon atoms Cartesian coordinates and by successive transposition. By this way an ensemble of shuffled random difference matrices is constructed, essentially as reported above for the empirical correlation matrix; we will refer to this as the shuffled ensemble. The other method that we use consists in the construction of a matrix as an N-dimensional Brownian process, having the same dimension and variance (along the columns) of the empirical matrix. After transposition, these matrices are treated as the empirical one, and we will refer to these matrices in the forthcoming discussion as the Brownian ensemble.

An interesting finding of the RMT is that, for each random matrix ensemble, a particular distribution of eigenvalues must be expected. This is the main reason for which RMT is extremely useful in signal extraction: these eigenvalue limiting densities indicate the cutoff between noise and signal. In the limit $N, T \to \infty$ and $Q = T/N \geq 1$ (the non-degenerate case) the eigenvalue density for the Wishart matrices is described by the Marčenko–Pastur distribution [35,36]:

$$\rho_W = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda} \tag{7}$$

$$\lambda_{min}^{max} = \sigma^2 \left(1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}}\right) \tag{8}$$

where $\lambda_{min} \leq \lambda \leq \lambda_{max}$. In case of large, but finite, N and T, blurred boundaries of this distribution are expected. This is described, for the ensemble we are interested in, by the Tracy–Widom distribution [37,38]. In the limit of large N, T, the Tracy–Widom distribution width is given by [37,33]:

$$\sigma_{TW} = \sqrt{1/Q} (\lambda_{max}/N)^{2/3}. \tag{9}$$

What was reported above permits to differentiate non-random components in the correlation matrix. We consider the correlation matrix as composed by two different components

$$C = C_r + C_{nr} \tag{10}$$

where C_r is the random component and C_{nr} the non-random one. This is reminiscent of the correlated Wishart matrices [33,39], a particular Wishart ensemble in which to a classical random Wishart matrix a perturbation is added. In this particular case, the λ_1 is repelled to the right from the rest of the spectrum, provided that the perturbation is above a certain threshold. On these basis, we obtain the random component of the correlation matrix using the following algorithm: *i*) calculate the eigenvalues and the eigenvectors of the correlation matrix C; *ii*) order the eigenvector matrix V as the rank of corresponding eigenvalues; *iii*) construct an all zeros matrix Z, of same size of C, except the element $(1, 1)$ which equals λ_{max}; and *iv*) calculate the matrix $U = VZ(V)^{-1}$ and calculate the new correlation matrix $C_n = C - U$.

We iterate this process until in the U matrix non-random elements are present. On the basis of the empirical correlation matrix eigenvalue distribution and the RMT expected distribution, we can estimate how many times the algorithm must be iterated until the smallest non-random eigenvalue above the Wishart threshold has been removed from the correlation matrix. The result of this iterative calculation is the C_r matrix, so we obtain the non-random correlation matrix by simple subtraction. Finally, the centered and scaled difference matrix X is projected, by performing the dot product, on the ordered eigenvectors obtained from the cleaned correlation matrix C_{nr}.

3. Results and discussion

Molecular dynamics simulations have been performed on the two-helix bundle protein Cox17. This is a small, 69 residues long, copper chaperone in mitochondria, characterized by a significant conformational flexibility [40,41]. Cox17 works as a metal chaperone on Sco1 and Sco2 [42], which in turn are involved during the copper insertion in the so-called Cu_A site of the cytochrome *c* oxidase [43]. Secondary structure elements of Cox17 are two α-helices (residues 27–39 and 48–60) and two coiled-coil regions; this coiled-coil-helix–coiled-coil-helix domain is a structural motif that was observed also in other mitochondrial proteins [26,40,41]. Our molecular dynamics set-up has been deliberately designed as a minimal one, in order to obtain what is generally considered to be a worst case, i.e., for the purpose of this report, a situation in which the projections on the first principal axes clearly show a high cosine content. This molecular system has been previously used to study the protein dynamics complexity in a minimal computational set-up [27]. It has been shown that, besides the α-hairpin, Cox17 exhibits low levels of secondary structure, being able to explore an unusually flat conformational landscape, similar to the intrinsically unstructured proteins [44,45], but anyway able to maintain a defined three-dimensional structure, at least in the nanosecond time-scale [27]. This behavior is well described by root mean square deviation analysis: for different simulations at 310 K, we observe values ranging from (± standard deviation) 3.3 ± 0.7 Å to 4.5 ± 1.0 Å for all the protein atoms (but excluding hydrogen atoms), corresponding to only 0.8 ± 0.14 Å and 0.9 ± 0.20 Å respectively, if it was calculated on the α-carbon atoms in the two-helix hairpin region. Even if in the following discussion we analyze only a single simulation run (to which we will refer as the empirical matrix), consider that rather identical results have been obtained for all the replicate simulations. As expected, because of the very short simulation time, PCA shows a typical cosine-shaped

appearance of low index projections on principal axes, as reported in Fig. 1.

We can also observe a particularly striking similarity between the projections on the first principal axes of the protein simulation matrix and of a random diffusion one (see Fig. 1). This last matrix represents a 207-dimensional Brownian dynamics simulation having, for each dimension, the same Gaussian distribution of the protein simulation matrix. This is the typical case described in Ref. [16] and theoretically justified therein. It should be recalled, however, that on the basis of the Karhunen–Loève theorem, a Wiener process, i.e. the continuous-time counterpart of Brownian motion, at least in some conditions, can be expressed as an infinite linear combination of sine-shaped functions, so the appearance of sine- (or cosine-) shaped functions in the PCA of a multidimensional Brownian motion is not particularly surprising. We suggest that, from a physical point of view, what this cosine-shaped appearance in the low-index principal projections means is just that the system experiences a barrier-free diffusion. It may be that barriers exist, but they are not encountered due to a too short-time observation (or allowed evolution) of the system; or they are too shallow to represent a significant kinetic barrier at the allowed energy (temperature); or simply they do not exist, as in the mathematical multidimensional Brownian diffusion. We cannot guess what the right option is by simply observing these cosine-shaped projections. Anyway, when the protein trajectory is projected on the first two principal axes, the half and full cosine produce a semi-circle that could be erroneously interpreted as the transition from a well-defined conformation to another (see also below).

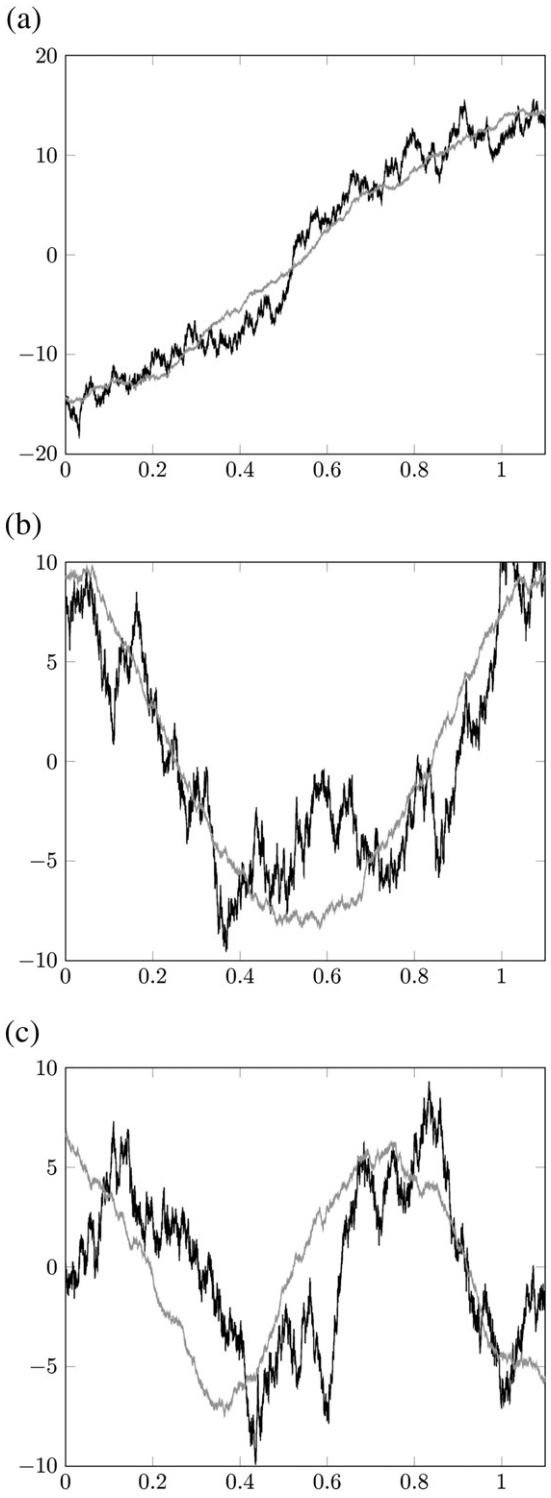


Fig. 1. Principal component analysis. First (a), second (b) and third (c) PC projections of the Cox17 empirical simulation matrix (black line) and of a 207-dimensional Brownian matrix (gray line) are reported. Note the striking similarities and the evident equivalent cosine content of the low index PCs for both matrices.

As stated above, cosine-like shaped projections on first principal components indicate that there were no barriers encountered by the system during its evolution. But does this mean that the system dynamics is featureless, exactly as a pure multidimensional Brownian dynamics? This is a question of practical as well as theoretical importance. Practical because, besides the presence or absence of barriers in the explored conformational landscape, one could be still interested in determining if clusters of conformations or coherent motions can be resolved in molecular dynamics simulations in a meaningful way. But the question is of theoretical importance also, because if, on a very short time-scale, protein dynamics could really be described as a pure Brownian dynamics, then, taking the argument to the limit, the whole protein dynamics is nothing but a type of, maybe special, Brownian dynamics. But this leads to a ballistic behavior on long time-scale, as pointed out by Hess [16,17]. To avoid this, it is necessary to postulate a mechanism through which non-Brownian behavior emerges from a patchwork of short Brownian pieces. But as we shall see below, this is not necessary because, even at very short times, protein dynamics is not really Brownian.

One of the most powerful tools for extraction of collective behavior from a sea of noise is the Random Matrix Theory (RMT). The study of matrices with random entries goes back as far as the 1920s, when it was introduced by Wishart [34] in the field of mathematical statistics. It was however in the 1950s that the wide use of random matrices has been introduced in physical modeling by Wigner [46,47] and Dyson [48]. Since then, RMT has been proven to be an astonishing successful technique, which has found a vast array of applications, ranging from the original quantum mechanics ones, to the complex systems analysis, including financial markets [33,49,50]. Surprisingly enough, besides few examples [51–53], no significant wide exploitation of RMT in protein dynamics analysis can be reported so far.

One critical point of the PCA-based analyses is a reliable empirical determination of the correlation matrix, which generally turns out to be a difficult task. For a set of N system degrees of freedom, the correlation matrix, which is symmetric by construction, contains $N(N-1)/2$ entries, obviously besides the diagonal ones. Each of these entries must be determined from N time series of length T, and if T is not very large compared to N, obtained covariance values are intrinsically noisy. In this case, the empirical correlation matrix is to a large extent random. From this point of view, it is important to develop methods able to distinguish true system signals from noise. As a null hypothesis, we compare the simulation correlation matrix properties to what is expected in the case of purely random matrices. Deviations from the random matrices' behavior could suggest the presence of useful information, i.e. the presence of non-random motion in the system under analysis. Here we use an empirical correlation matrix constructed from the degrees of freedom change time series, $\Delta x_i(t)$ (where i labels the degree

of freedom and t the time) in the unit time. This last is related to the sampling frequency (which is, in the reported empirical matrix, 250 fs^{-1}). From a physical viewpoint, these quantities correspond to velocity. However, the displacement matrix, which is more often employed in molecular dynamics PCA calculations, can be used too in RMT based analysis.

If we consider the eigenspectrum of the empirical correlation matrices, the essential problem to overcome is to distinguish between the correlations originating from real dependences among the degrees of freedom and the false (or random) correlations which are simply a consequence of the finite length of signals. For this purpose, RMT can be used, because it offers universal forecasts for specific matrix ensembles. In the case of the empirical correlation matrices obtained from molecular dynamics experiments of dimension $N \times T$ (where N represents the number of Cartesian coordinates and T the observed time points), the relevant one is the Wishart matrix ensemble (see Methods section). In the limit $N, T \to \infty$ the eigenvalue density for the Wishart matrices is given by the Marčenko–Pastur distribution. In Fig. 2 the distribution of eigenvalues from the empirical correlation matrix is reported (panel a).

If we compare this eigenvalue distribution with those obtained from the shuffled simulation matrix ensemble (Fig. 2, panel b) and with the eigenvalue distribution obtained from a 207-dimensional Brownian matrix ensemble (Fig. 2, panel c), we can easily see that an enormous difference is evident. The eigenvalue distributions in the shuffled and Brownian cases are clearly different from the empirical one. The shuffled matrix and the Brownian ensembles behave exactly as predicted by the RMT. The best-fitting Marčenko–Pastur distribution for the reported shuffled ensemble has λ_{min} and λ_{max} values of 0.52 and 1.62, respectively, and also the fitted Q value is well in agreement with the expected one. The Q value calculated by best-fitting the Marčenko–Pastur distribution for the shuffled ensemble is 21.05, when the expected (based on the matrix dimensions T and N, because Q is the T/N ratio) was 21.25. On the basis of fitted parameter, for the shuffled ensemble we estimate a value of 0.009 for the width of the Tracy–Widom distribution. For the Brownian matrix ensemble, the best-fitting Marčenko–Pastur distribution shows a λ_{min}, λ_{max} and Q of 0.60, 1.51 and 23.0, respectively, with a calculated Tracy–Widom distribution width of 0.008. Even considering the expected λ_{max} smearing, several eigenvalues of the empirical correlation matrix under analysis absolutely cannot be considered as random. This is a crucial point in our analysis: also in presence of cosine-shaped projections in low-index principal axes, the system is completely different from the Brownian ensemble, as well as from the random shuffled one. It is important to stress here that this type of result is absolutely not to be considered an artifact due to the particular system under consideration. The RMT can be used to determine the random eigenvalue boundary in all datasets that are obtained by molecular dynamics; an example of this type of analysis is shown in Supplementary Fig. 1, where the results obtained from a simulation carried out as described in Ref. [54] are shown. Even if in the forthcoming analysis we will focus our attention on the largest eigenvalues deviating from RMT predictions, it should be noted that the spectrum of the empirical correlation matrix deviates from random matrices' behavior for smallest eigenvalues too. In the case of empirical correlation matrices, eigenvalues below the RMT allowed ones can be associated to eigenvectors describing correlations, or anticorrelations, between pairs or small groups of degrees of freedom. We will not discuss them further because they represent very local, instead of global, system motions. However, in some applications (for example in the reconstruction of residue connections in dynamically relevant networks) a detailed analysis of these non-random and small eigenvalues associated eigenvectors could be useful.

Once we have determined that the eigenvalue distribution of the simulation matrix behaves differently from what is predicted by the RMT, where its random counterparts perfectly obey to this theory, it is interesting to test how long the simulation must be in order to observe eigenvalues, with associated eigenvectors, out of the random range. We

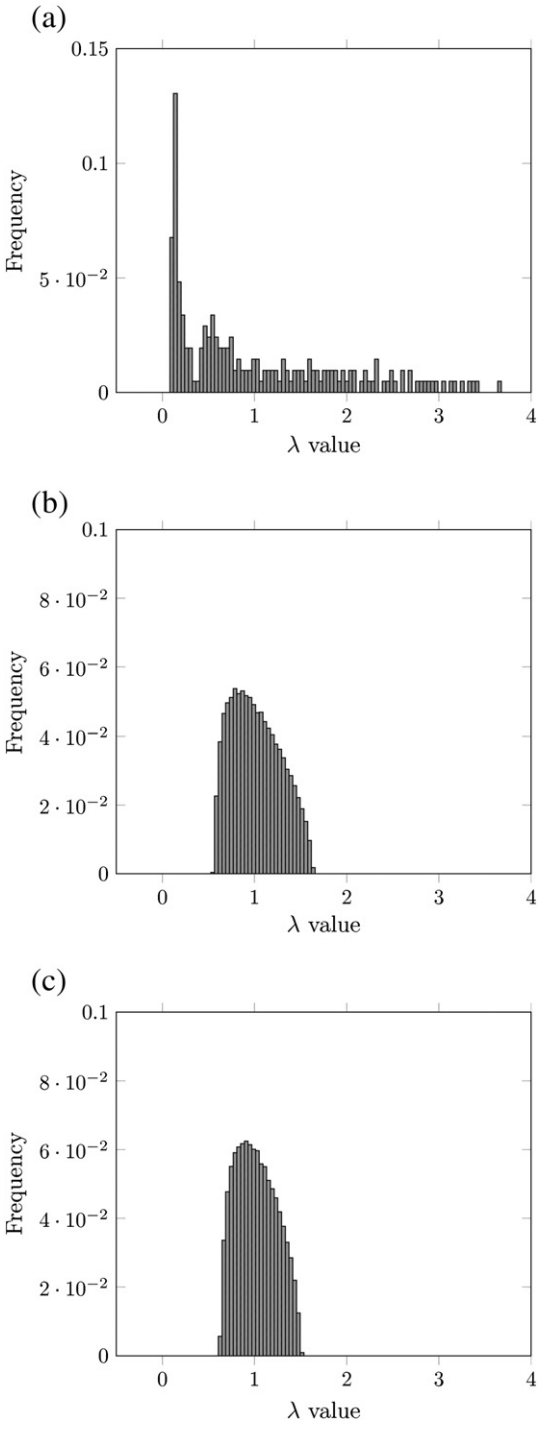


Fig. 2. Spectral analysis. Spectra of the Cox17 empirical correlation matrix (a) and of the shuffled (b) and Brownian (c) ensembles are reported. Each ensemble was composed of 100 different shuffled and Brownian matrices, respectively; spectra were calculated as reported in the Methods section.

perform such analysis by considering a series of shortened matrices, that were built by selecting from the original simulation matrix only the rows corresponding to time from 0 to $t_c \leq T$. By this way, we can observe the spectral evolution of the simulation at several Q values. For each of these partial matrices we calculate the expected λ_{max} from the Marčenko–Pastur distribution and the Tracy–Widom distribution width, σ_{TW}. From these values we can obtain, from the spectra of the shortened matrices, the number of eigenvalues above the random range. The results are reported in Table 1. Columns in this table represents, respectively: the Q value obtained from the matrix dimensions; the RMT predicted λ_{max};

Table 1
RMT predicted parameters and number of eigenvalues above the random matrix range at different Q values.

Q	λ_{max}	σ_{TW}	$\lambda > \lambda_{max}$	$\lambda > \lambda_{max}$ ($p \leq 0.001$)
1.01	3.8050	0.0665	13	10
5.0	2.0390	0.0209	35	32
10.0	1.7318	0.0130	46	45
15.0	1.5827	0.0100	53	51
21.25	1.4806	0.0081	57	57

the expected Tracy–Widom distribution width σ_{TW}; the number of eigenvalues above the RMT predicted λ_{max} ($\lambda > \lambda_{max}$); the number of eigenvalues above the RMT threshold with $p \leq 0.001$ (the rightmost column). The significance levels are calculated as in Patterson et al. [55].

An amazing finding of this analysis is that, in this simulation, a significant number of eigenvalues are above the predicted RMT threshold even in the case of $Q = 1.01$. At this Q value we are considering a situation near the limit of the degenerate case, but still 10 eigenvalues are out of the random matrix range with $p \leq 0.001$. Similar results are obtained using the shuffled (or the Brownian) ensembles at different Q values in order to assess the random eigenvalue boundaries, as reported in Fig. 3. It should be noted that the exact number of eigenvalues above the random threshold varies from one simulation to another, particularly at close to unit Q values.

Moreover, the non-random eigenvalue number oscillates if we analyze different time subsets within a simulation. Using a sliding window protocol set to $Q = 1.01$ over all the Cox17 simulations performed in this work, and considering as non-random eigenvalues with $p \leq 0.001$ those calculated as reported above [55], we have an estimate of the non-random eigenvalue number oscillation within simulations. The minimum non-random eigenvalue number was 9, the maximum was 13, with mean, median and mode equal to 11 (considering an ensemble of more than 8000 time subsets). The estimate number of non-random eigenvalues depends also from the threshold calculation method, again particularly at Q values near the degenerate case. What must be highlighted here is the fact that, for all $Q > 1$, there are some eigenvalues above the random threshold, howsoever calculated. As can be expected, the number of non-random large eigenvalues increases with increasing Q; this is a consequence of the intuitive fact that more precise measurements (i.e. longer simulations) allow to better distinguish between random motion and molecular movement. In long simulations the number of eigenvalues above the Wishart range oscillates around a limiting value, which suggests a method for estimating the so-called convergence of a simulation (not shown).

In order to eliminate the random contributions in the covariance matrix, we iteratively construct a correlation matrix containing only components above those predicted by RMT, as detailed in the Methods section. In Fig. 4 the correlation matrices before and after this cleaning procedure are reported.

Consider that after this iteration, the original correlation matrix (Fig. 4, lower left triangle) has been separated in a non-random one (Fig. 4, upper right triangle) and a matrix containing only random entries (not shown). We use the eigenvectors of the non-random part of the correlation matrix for the subsequent eigenvector analysis of the empirical matrix. As usually done in PCA, by this way we obtain a rotation of the initial dataset. By projecting the empirical matrix on the non-random eigenvectors we lose part of the original variance, but the global motions that we can now observe are surely out of the random range. Correlations among distant residues in this cleaned matrix representation appear more evident (although faint) than in the original, noise-blurred, one. It should be noted that a similar iterative cleaning on the Brownian correlation matrices leads to a complete loss of the correlated signals originally present, simply represented by the correlation of each element with itself. This is due to the fact that, in this case, a large, or better a major, part of the original variance is still in the residual correlation matrix. The residual correlation matrix corresponds to the random one in the case of the empirical correlation matrix iteration.

The appearance of low index projections on the non-random basis vectors is, obviously, still cosine-like (not shown), but this is simply due to the fact that our iterative calculation selects eigenvectors from the original correlation matrix. Moreover, this is in agreement with the observation that protein dynamics inside a potential well approaches a Brownian-like regime on the whole protein scale [27]. Frequency spectra analysis on molecular dynamics time-series has led to the proposal that the observed in-time correlations can be described mathematically as the convolution of a Gaussian signal with an n-body decay term. This last term is due to the networked nature of the whole protein system [27].

A striking difference that can be observed when one performs a spectral analysis of molecular dynamics-deriving matrices from one hand and the pure Brownian (or also shuffled) matrices on the other is related to the eigenvector 'structures'. In the latter case no meaningful patterns can be observed at all, but in the former case, i.e. in the molecular dynamics deriving matrices, non-random eigenvectors show interesting patterns, for which a meaningful interpretation can be made. Moreover, eigenvector component distributions for the correlation matrices of the Brownian and shuffled ensembles fall in the predicted Porter–Thomas distribution for random matrices [33], but in the case of the empirical correlation matrices this is true only for eigenvectors in the Wishart range (not shown). In the reported simulation, non-random eigenvector analysis reveals interesting rate patterns associated to secondary structure motions such as α-helix bending, stretching and displacements. For example, in Fig. 5 the rate distribution derived from the first eigenvector is shown. The associated eigenvalue is, as demonstrated above, out of the Wishart range with absolute confidence. It should be emphasized that this result is not important because we observe a *particular* type of motion in our model system. Instead it is important because we observe a non-random and coherent protein motion where a random, Brownian type, dynamics was expected (high cosine content dynamics).

This means that, even in the case of very short molecular dynamics, non-random events can be easily discriminated from random ones. This discrimination occurs in the presence of clearly cosine-shaped projections on the firsts PCs. This leads us to conclude that the short-time dynamics of Cox17 is Brownian-mimetic but not really Brownian. This can be easily extended also to other protein systems, because our reported example on the short-time dynamics of Cox17 cannot be considered as an exception. It is the underlying structure of the correlation matrix, expressed by the presence of evocative non-random eigenvectors, that differentiates the protein dynamics from the pure Brownian motion.

4. Conclusions

This work shows that RMT can be of wide use in molecular dynamics analysis, because it permits to accurately discriminate between random components of motion from non-random ones. True correlations can be rigorously discriminated from those which are derived from the, always, finite length of simulation. If we are not in presence of a true Brownian system, but instead of a correlated one, we can still extract useful information on the system out of the sea of noise, and the non-random eigenvectors could be related to specific system features in a meaningful way. As shown above, the boundary of non-random eigenvalues, and the associated eigenvectors, can be estimated from only a few easily calculable parameters. Knowing Q, which is determined by the simulation time and the protein dimension, and knowing the whole simulation matrix variance, we can compare the simulation matrix spectrum with the expected Wishart boundary at a very low computational cost. The reported analysis shows that this works even in the presence of cosine-shaped projections on the low-index principal components. Indeed we have shown that there are substantial differences between true Brownian systems and, even weakly, correlated

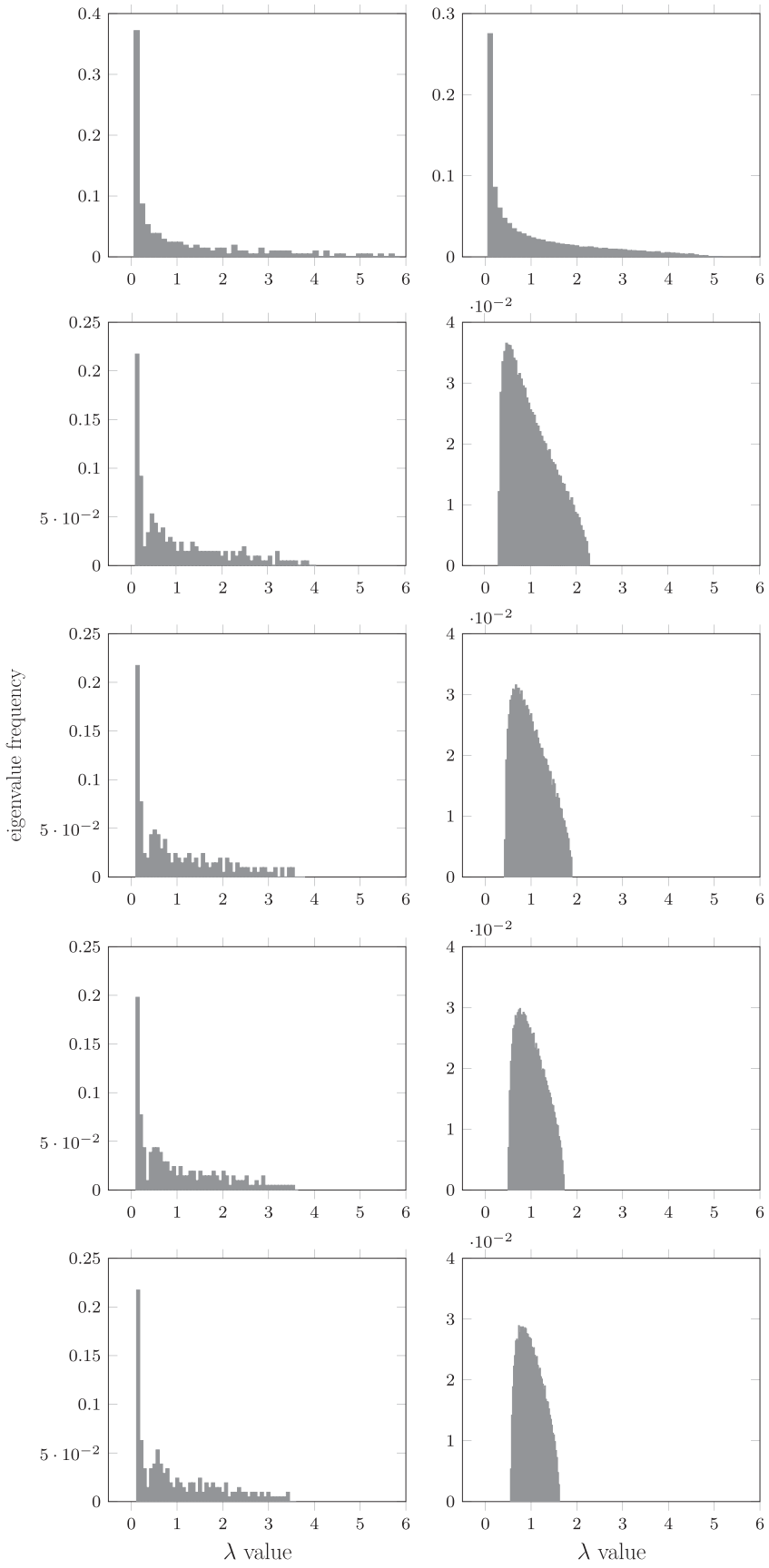


Fig. 3. Spectral analysis at different *Q* values. Spectra of the Cox17 empirical correlation matrices (left panels) and of the relative shuffled ensembles (right panels) are reported. Each shuffled ensemble was composed of 100 different matrices. Showed *Q* values are, from top to bottom, 1.1, 6.1, 11.1, 16.1 and 21.1, respectively. Spectra were calculated as reported in the Methods section.

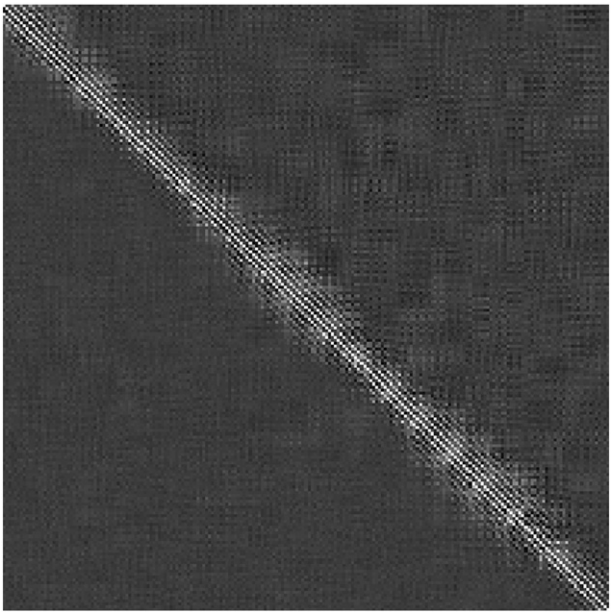


Fig. 4. Correlation matrices. The original correlation matrix C (lower left) and the non-random part C_{nr} (upper right) of Cox17 simulation are reported as image. Scale ranges from dark gray to white, corresponding respectively to − 0.2 and 0.6. Consider that the minimum and maximum observed numerical entries in these matrices were − 0.3618 and 0.9998 respectively.

ones. The presence of cosine-shaped trajectories, obtained by performing the PCA on the protein dynamics, simply means that a barrier-free evolution has been experienced by the system; this observation can also be extended to other systems characterized by a large degree of freedom number. We cannot say nothing else more from this, because the question on the presence or absence of barriers in the system dynamical landscape is not resolvable by this way. Anyway, our RMT based analysis offers a simple solution to the, conceptually discomforting, paradox of the apparent similarity between polymer dynamics inside a potential well and the pure mathematical Brownian dynamics.

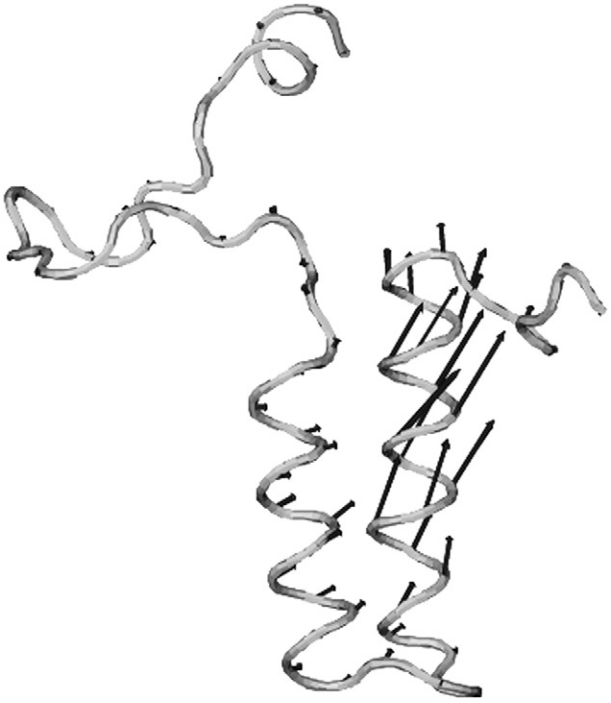


Fig. 5. Eigenvectors' analysis. Cox17 structure (light gray) and α-carbon atom rate vectors (dark gray) are reported. Drawn vectors are proportional to the calculated rate from the first eigenvector (relative to the whole simulation), which is with absolute confidence out of the Wishart range. This eigenvector correspond essentially to a stretching motion of the C-terminal α-helix.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.bpc.2014.08.007.

Acknowledgments

The author thanks Fabrizio Bossis for assistance in molecular dynamics experiments. The author also wishes to thank the referees and the Editor for their insightful advice.

References

[1] M. Karplus, J.A. McCammon, Molecular dynamics simulations of biomolecules, Nat. Struct. Mol. Biol. 9 (9) (2002) 646–652, http://dx.doi.org/10.1038/nsb0902-646.
[2] F. Bossis, L.L. Palese, Molecular dynamics in cytochrome *c* oxidase Mössbauer spectra deconvolution, Biochem. Biophys. Res. Commun. 404 (1) (2011) 438–442, http://dx.doi.org/10.1016/j.bbrc.2010.11.140.
[3] I. Daidone, A. Amadei, Essential dynamics: foundation and applications, WIREs, Comput. Mol. Sci. 2 (5) (2012) 762–770, http://dx.doi.org/10.1002/wcms.1099.
[4] A. Kitao, N. Go, Investigating protein dynamics in collective coordinate space, Curr. Opin. Struct. Biol. 9 (2) (1999) 164–169, http://dx.doi.org/10.1016/S0959-440X(99)80023-2.
[5] B. Brooks, M. Karplus, Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor, Proc. Natl. Acad. Sci. U. S. A. 80 (21) (1983) 6571–6575.
[6] M. Levitt, C. Sander, P.S. Stern, The normal modes of a protein: native bovine pancreatic trypsin inhibitor, Int. J. Quantum Chem. 24 (S10) (1983) 181–199, http://dx.doi.org/10.1002/qua.560240721.
[7] M. Karplus, J.N. Kushick, Method for estimating the configurational entropy of macromolecules, Macromolecules 14 (2) (1981) 325–332, http://dx.doi.org/10.1021/ma50003a019.
[8] T. Ichiye, M. Karplus, Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations, Proteins 11 (3) (1991) 205–217, http://dx.doi.org/10.1002/prot.340110305.
[9] R. Levy, A. Srinivasan, W. Olson, J. McCammon, Quasi-harmonic method for studying very low frequency modes in proteins, Biopolymers 23 (6) (1984) 1099–1112, http://dx.doi.org/10.1002/bip.360230610.
[10] A. Amadei, A. Linssen, H.J. Berendsen, Essential dynamics of proteins, Proteins 17 (4) (1993) 412–425, http://dx.doi.org/10.1002/prot.340170408.
[11] R. Saegusa, H. Sakano, S. Hashimoto, Nonlinear principal component analysis to preserve the order of principal components, Neurocomputing 61 (2004) 57–70, http://dx.doi.org/10.1016/j.neucom.2004.03.004.
[12] P.H. Nguyen, Complexity of free energy landscapes of peptides revealed by nonlinear principal component analysis, Proteins 65 (4) (2006) 898–913, http://dx.doi.org/10.1002/prot.21185.
[13] O.F. Lange, H. Grubmüller, Full correlation analysis of conformational protein dynamics, Proteins 70 (4) (2008) 1294–1312, http://dx.doi.org/10.1002/prot.21618.
[14] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323, http://dx.doi.org/10.1126/science.290.5500.2319.
[15] P. Das, M. Moll, H. Stamati, L.E. Kavraki, C. Clementi, Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction, Proc. Natl. Acad. Sci. U. S. A. 103 (26) (2006) 9885–9890, http://dx.doi.org/10.1073/pnas.0603553103.
[16] B. Hess, Similarities between principal components of protein dynamics and random diffusion, Phys. Rev. E. 62 (6) (2000) 8438, http://dx.doi.org/10.1103/PhysRevE.62.8438.
[17] B. Hess, Convergence of sampling in protein simulations, Phys. Rev. E. 65 (3) (2002) 031910, http://dx.doi.org/10.1103/PhysRevE.65.031910.
[18] J.B. Clarage, T. Romo, B.K. Andrews, B.M. Pettitt, G.N. Phillips, A sampling problem in molecular dynamics simulations of macromolecules, Proc. Natl. Acad. Sci. U. S. A. 92 (8) (1995) 3288–3292, http://dx.doi.org/10.1073/pnas.92.8.3288.
[19] B. De Groot, D. van Aalten, A. Amadei, H. Berendsen, The consistency of large concerted motions in proteins in molecular dynamics simulations, Biophys. J. 71 (4) (1996) 1707–1713, http://dx.doi.org/10.1016/S0006-3495(96)79372-4.
[20] A. Amadei, M.A. Ceruso, A. Di Nola, On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations, Proteins 36 (4) (1999) 419–424.
[21] O.F. Lange, H. Grubmüller, Can principal components yield a dimension reduced description of protein dynamics on long time scales? J. Phys. Chem. B 110 (45) (2006) 22842–22852, http://dx.doi.org/10.1021/jp062548j.
[22] M.A. Balsera, W. Wriggers, Y. Oono, K. Schulten, Principal component analysis and long time protein dynamics, J. Phys. Chem. 100 (7) (1996) 2567–2572, http://dx.doi.org/10.1021/jp9536920.
[23] P. Diaconis, Patterns in eigenvalues: the 70th Josiah Willard Gibbs lecture, Bull. Am. Math. Soc. 40 (2) (2003) 155–178, http://dx.doi.org/10.1090/S0273-0979-03-00975-3.
[24] A. Edelman, N.R. Rao, Random matrix theory, Acta Numerica 14 (1) (2005) 233–297, http://dx.doi.org/10.1017/S0962492904000236.
[25] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242, http://dx.doi.org/10.1093/nar/28.1.235.

[26] C. Abajian, L.A. Yatsunyk, B.E. Ramirez, A.C. Rosenzweig, Yeast cox17 solution structure and copper (i) binding, J. Biol. Chem. 279 (51) (2004) 53584–53592, http://dx.doi.org/10.1074/jbc.M408099200.
[27] L.L. Palese, Protein dynamics: complex by itself, Complexity 18 (3) (2013) 48–56, http://dx.doi.org/10.1002/cplx.21434.
[28] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, K. Schulten, Namd2: greater scalability for parallel molecular dynamics, J. Comput. Phys. 151 (1) (1999) 283–312, http://dx.doi.org/10.1006/jcph.1999.6201.
[29] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, K. Schulten, Scalable molecular dynamics with namd, J. Comput. Chem. 26 (16) (2005) 1781–1802, http://dx.doi.org/10.1002/jcc.20289.
[30] A.D. MacKerell, D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wirkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins, J. Phys. Chem. B 102 (18) (1998) 3586–3616, http://dx.doi.org/10.1021/jp973084f.
[31] A.D. MacKerell, M. Feig, C.L. Brooks, Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, J. Comput. Chem. 25 (11) (2004) 1400–1415, http://dx.doi.org/10.1002/jcc.20065.
[32] W. Humphrey, A. Dalke, K. Schulten, Vmd: visual molecular dynamics, J. Mol. Graph. 14 (1) (1996) 33–38, http://dx.doi.org/10.1016/0263-7855(96)00018-5.
[33] J. Kwapień, S. Drożdż, Physical approach to complex systems, Phys. Rep. 515 (3) (2012) 115–226, http://dx.doi.org/10.1016/j.physrep.2012.01.007.
[34] J. Wishart, The generalised product moment distribution in samples from a normal multivariate population, Biometrika 20 (1/2) (1928) 32–52.
[35] V.A. Marčenko, L.A. Pastur, Distribution of eigenvalues for some sets of random matrices, Math. Sb. 1 (4) (1967) 457–483, http://dx.doi.org/10.1070/SM1967v001n04ABEH001994.
[36] A.M. Sengupta, P.P. Mitra, Distributions of singular values for some random matrices, Phys. Rev. E. 60 (3) (1999) 3389, http://dx.doi.org/10.1103/PhysRevE.60.3389.
[37] C.A. Tracy, H. Widom, Level-spacing distributions and the airy kernel, Commun. Math. Phys. 159 (1) (1994) 151–174, http://dx.doi.org/10.1007/BF02100489.
[38] C.A. Tracy, H. Widom, Distribution functions for largest eigenvalues and their applications, arXiv, preprint math-ph/0210034, 2002.
[39] K.E. Bassler, P.J. Forrester, N.E. Frankel, Eigenvalue separation in some random matrix models, J. Math. Phys. 50 (2009) 033302, http://dx.doi.org/10.1063/1.3081391.
[40] F. Arnesano, E. Balatri, L. Banci, I. Bertini, D.R. Winge, Folding studies of cox17 reveal an important interplay of cysteine oxidation and copper binding, Structure 13 (5) (2005) 713–722, http://dx.doi.org/10.1016/j.str.2005.02.015.
[41] L. Banci, I. Bertini, S. Ciofi-Baffoni, A. Janicka, M. Martinelli, H. Kozlowski, P. Palumaa, A structural–dynamical characterization of human cox17, J. Biol. Chem. 283 (12) (2008) 7912–7920, http://dx.doi.org/10.1074/jbc.M708016200.
[42] P.A. Cobine, F. Pierrel, D.R. Winge, Copper trafficking to the mitochondrion and assembly of copper metalloenzymes, Biochim. Biophys. Acta 1763 (7) (2006) 759–772, http://dx.doi.org/10.1016/j.bbamcr.2006.03.002.
[43] S. Papa, N. Capitanio, G. Capitanio, L.L. Palese, Protonmotive cooperativity in cytochrome *c* oxidase, Biochim. Biophys. Acta 1658 (1) (2004) 95–105, http://dx.doi.org/10.1016/j.bbabio.2004.04.014.
[44] T. Mittag, L.E. Kay, J.D. Forman-Kay, Protein dynamics and conformational disorder in molecular recognition, J. Mol. Recognit. 23 (2) (2010) 105–116, http://dx.doi.org/10.1002/jmr.961.
[45] H.J. Dyson, P.E. Wright, Intrinsically unstructured proteins and their functions, Nat. Rev. Mol. Cell Biol. 6 (3) (2005) 197–208, http://dx.doi.org/10.1038/nrm1589.
[46] E.P. Wigner, Characteristic vectors of bordered matrices with infinite dimensions, Ann. Math. 62 (3) (1955) 548–564.
[47] E.P. Wigner, Random matrices in physics, SIAM Rev. 9 (1) (1967) 1–23, http://dx.doi.org/10.1137/1009001.
[48] F.J. Dyson, Statistical theory of the energy levels of complex systems. I, J. Math. Phys. 3 (1) (1962) 140–156, http://dx.doi.org/10.1063/1.1703773.
[49] L. Laloux, P. Cizeau, J.-P. Bouchaud, M. Potters, Noise dressing of financial correlation matrices, Phys. Rev. Lett. 83 (7) (1999) 1467, http://dx.doi.org/10.1103/PhysRevLett.83.1467.
[50] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A. Nunes Amaral, H.E. Stanley, Universal and nonuniversal properties of cross correlations in financial time series, Phys. Rev. Lett. 83 (7) (1999) 1471, http://dx.doi.org/10.1103/PhysRevLett.83.1471.
[51] R. Potestio, F. Caccioli, P. Vivo, Random matrix approach to collective behavior and bulk universality in protein dynamics, Phys. Rev. Lett. 103 (26) (2009) 268101, http://dx.doi.org/10.1103/PhysRevLett.103.268101.
[52] Y. Matsunaga, S. Fuchigami, A. Kidera, Multivariate frequency domain analysis of protein dynamics, J. Chem. Phys. 130 (2009) 124104, http://dx.doi.org/10.1063/1.3090812.
[53] M. Yamanaka, Random matrix theory analysis of cross correlations in molecular dynamics simulations of macro-biomolecules, J. Phys. Soc. Jpn. 82 (8) (2013) 083801, http://dx.doi.org/10.7566/JPSJ.82.083801.
[54] F. Bossis, L.L. Palese, Amyloid beta (1–42) in aqueous environments: effects of ionic strength and e22q (Dutch) mutation, Biochim. Biophys. Acta 1834 (12) (2013) 2486–2493, http://dx.doi.org/10.1016/j.bbapap.2013.08.010.
[55] N. Patterson, A.L. Price, D. Reich, Population structure and eigenanalysis, PLoS Genet. 2 (12) (2006) e190, http://dx.doi.org/10.1371/journal.pgen.0020190.