Generalized Correlation for Biomolecular Dynamics

Oliver F. Lange and Helmut Grubmüller*
Department of Theoretical and Computational Biophysics, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

ABSTRACT **Correlated motions in biomolecules are often essential for their function, e.g., allosteric signal transduction or mechanical/thermodynamic energy transport. Because correlated motions in biomolecules remain difficult to access experimentally, molecular dynamics (MD) simulations are particular useful for their analysis. The established method to quantify correlations from MD simulations via calculation of the covariance matrix, however, is restricted to linear correlations and therefore misses part of the correlations in the atomic fluctuations. Herein, we propose a general statistical mechanics approach to detect and quantify any correlated motion from MD trajectories. This generalized correlation measure is contrasted with correlations obtained from covariance matrices for the B1 domain of protein G and T4 lysozyme. The new method successfully quantifies correlations and provides a valuable global overview over the functionally relevant collective motions of lysozyme. In particular, correlated motions of helix 1 together with the two main lobes of lysozyme are detected, which are not seen by the conventional covariance matrix. Overall, the established method misses more than 50% of the correlation. This failure is attributed to both, an interfering and unnecessary dependence on mutual orientations of the atomic fluctuations and, to a lesser extent, attributed to nonlinear correlations. Our generalized correlation measure overcomes these problems and, moreover, allows for an improved understanding of the conformational dynamics by separating linear and nonlinear contributions of the correlation. Proteins 2006;62:1053–1061.** © 2005 Wiley-Liss, Inc.

Key words: mutual information; conformational entropy; MD simulation; correlated motion; collective motion; protein dynamics; conformational dynamics; lysozyme; kullback-leibler divergence

INTRODUCTION

Correlated motions in biomolecules, in particular proteins, are ubiquitous and often essential for biomolecular function.[1] Examples are allosteric signal transduction, as in G protein coupled receptors,[2] or mechanical/thermodynamic energy transport, as in F_0/F_1-adenosine triphosphatase.[3] Furthermore, the energetics of protein function is often dominated by entropic contributions, which are directly linked to correlated atomic motion.[4–6] Correct assessment of correlated motions, both experimentally and from theory and simulations, is therefore crucial for a quantitative understanding of biomolecular function. The accurate characterization of correlated motions would also improve the interpretation of nuclear magnetic resonance (NMR) experiments[7] and X-ray diffusive scattering data.[8]

Recently, NMR relaxation experiments were proposed to directly probe correlated motions.[9] A subsequent computational study raised serious doubts about the validity of the interpretation of the obtained experimental results as correlated motion, however.[10] This work inspired the development of a generalized and nonlinear correlation measure, which allows the correct and complete assessment of correlated motion from molecular dynamics (MD) simulations.

The established method to quantify correlations from MD simulations, in analogy to the *Pearson correlation coefficient* rests on calculation of the normalized covariance matrix of atomic fluctuations, $C_{ij} = \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle / (\langle \mathbf{x}_i^2 \rangle \langle \mathbf{x}_j^2 \rangle)^{1/2}$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the positional fluctuation vectors of atoms i and j, respectively, in the molecular fixed frame.[11,12] As will be shown in the theoretical part, this established approach, however, misses a considerable fraction of the correlated motions and, therefore, usually underestimates atomic correlations. This limitation is mainly the result of two assumptions.

First, estimates of correlations from the Pearson coefficient are only strictly valid if $\mathbf{x}_i$ and $\mathbf{x}_j$ are colinear vectors, as already pointed out by Ichiye and Karplus.[12] Improved results are obtained with the method of *canonical* correlations[13] by choosing so-called *canonical* variables which furnish average colinearity, i.e., for every pair of atoms a different coordinate transformation is applied. In contrast to the Pearson correlation coefficient, canonical correlations do not differentiate between correlated and anticorrelated, i.e., *positively correlated*, and *negatively correlated* motion. Such a distinction becomes problematic in the multidimensional case, and thus has to be dropped for *any* meaningful correlation measure. Consider, e.g., two atoms that oscillate perfectly correlated in parallel directions. If the oscillation direction of one atom is rotated until both

Grant sponsor: Volkswagen foundation; Grant numbers: I/80436 and I/80585.

*Correspondence to: Helmut Grubmüller, Department of Theoretical and Computational Biophysics, Max-Planck-Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany. E-mail: hgrubmu@gwdg.de

Received 10 June 2005; Revised 15 August 2005; Accepted 26 August 2005

Published online 14 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20784

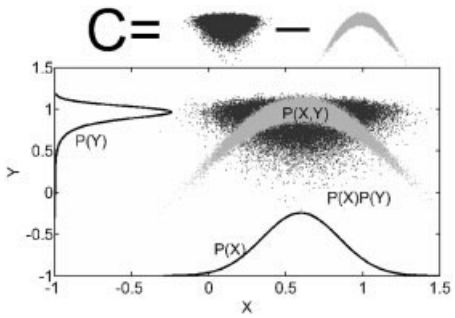

Fig. 1. Correlations of random variables are defined as deviation of their probability distribution (gray) from the hypothetical probability distribution of the independent random variables (black). In the sketch, the correlation between variables X and Y (gray in the scatter plot) is to be quantified. From the marginal distributions $P(X)$ and $P(Y)$ (black curves), one computes the hypothetical joint distribution for independent variables (black points) $P(X)P(Y)$. The difference between the given joint distribution and the hypothetical uncorrelated joint distribution yields the correlation measure C, as illustrated at the top of the graphic.

atoms oscillate antiparallel, the Pearson correlation coefficient changes from 1 to -1 and therefore has to cross zero, usually after rotation by 90°, i.e., when the directions are perpendicular. In this case, the vanishing correlation coefficient is highly misleading, because the motion of the two atoms is still perfectly correlated.

Second, use of the covariance matrix implies a Gaussian approximation of the underlying configurational space density. Therefore, this approach treats correlations in a quasi-harmonic, i.e., linear, approximation. Thus, the Pearson correlation coefficient, as well as the canonical correlation method, miss nonlinear correlations. Higher moment corrections are conceivable, but notoriously suffer from dramatic combinatorial increase of computational effort, and slow convergence, which renders the treatment of large systems such as proteins impossible.

As an efficient alternative, we propose here a general approach to quantify any correlated motion.

The proposed generalized correlation measure rests on the fundamental definition of independence of random variables. Accordingly, two random variables are independent, if and only if their joint distribution is a product of their marginal distributions, $P(X,Y) = P(X)P(Y)$. The basic idea is to quantify the correlation between variables X,Y as the deviation between both sides of the above equation, i.e., by the deviation from the case of two independent random variables (Fig. 1). As will be shown in Materials and Methods, this definition is equivalent to defining a correlation C as the well-known (Shannon) mutual information (MI),[14] $C[X,Y] = H[X] + H[Y] - H[X,Y]$, where H denotes the entropy of the random variables. This definition rests on the well-known inequality $H[X,Y] \leq H[X] + H[Y]$, which becomes an equality if and only if both variables are independent. This formulation is equivalent to an infinite moment expansion. Truncation at second moments yields a linearized MI which will be defined in Materials and Methods.

In this section, we also will review the definition of the Pearson correlation coefficient and its canonical interpretation. After this we recall basic properties of MI and propose to define the *generalized correlation coefficient* which maps the MI with values in the range $0,\ldots,\infty$ onto the more convenient interval [0,1] to allow a direct comparison with the Pearson correlation coefficient.

The impact of the known[12] problems of the Pearson correlation coefficient seems largely underrated, and the canonical correlation approach[13] is generally not applied. Herein, we quantify the inconsistencies and shortcomings of the Pearson correlation coefficient when applied to protein dynamics. To this end, two examples are studied, the B1 domain of protein G and T4 lysozyme (T4L). Using these examples, we will also show that our generalized correlation measure does not suffer from these shortcomings and, therefore, provides an accurate and complete quantification of correlations in protein dynamics.

MATERIALS AND METHODS

At first we introduce some notation. In this article, we focus on correlations of atomic fluctuations, i.e., of vectors in three-dimensional space. However, at some points, it is necessary to discuss correlations between one-dimensional variables. Therefore, we use the following notation. All positions of atoms (or other variables) are denoted by a vector $\mathbf{r} = (\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_N)$ with N components $\mathbf{r}_i \in \mathbb{R}^d$, with $d = 3$ for atoms. We refer to positional fluctuations, i.e., the deviation from the mean, $\mathbf{x} = \mathbf{r} - \langle\mathbf{r}\rangle$, with

$$\mathbf{x} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N) = (x_1^{(1)},x_1^{(2)},\ldots,x_1^{(d)},\ldots,x_N^{(1)},\ldots,x_N^{(d)})$$

and $\langle . \rangle$ denoting the ensemble average. With $p(\mathbf{x})$, we denote the corresponding probability density, which in the context of biomolecular dynamics is the canonical ensemble density $p(\mathbf{x}) = Z^{-1}\exp[-\beta V(\mathbf{x} + \langle\mathbf{r}\rangle)]$, where Z is the partition function, β the inverse temperature, and V the potential energy. Furthermore, we denote the marginal probability density by $p_i(\mathbf{x}_i) = \int p(\mathbf{x}) d\mathbf{x}_{j\neq i}$.

Pearson Correlation Coefficient

The established and intuitive method[11,12] to quantify the correlation between pairs of components (i,j) of the fluctuation vector $\mathbf{x}$ is

$$r[\mathbf{x}_i,\mathbf{x}_j] = \langle\mathbf{x}_i \cdot \mathbf{x}_j\rangle / (\langle\mathbf{x}_i^2\rangle\langle\mathbf{x}_j^2\rangle)^{1/2} \quad (1)$$

where the square brackets indicate the dependence on the whole ensemble of $\mathbf{x}_i,\mathbf{x}_j$.

In the one-dimensional case, r is called the *Pearson coefficient*, and it has a very straightforward and fairly general interpretation: Under the assumption that at least one variable is normally distributed, it yields the *coefficient of nondetermination*,

$$1 - r^2 = \frac{\langle (x_j - f(x_i))^2\rangle}{\langle x_j^2\rangle}, \quad (2)$$

of the best *linear* fit $f(x_i)$ to x_j. For the multidimensional case, an analogous interpretation of the Pearson coefficient is possible provided that the atoms (i,j) have unit variance $\langle x_i^{(k)} x_i^{(l)}\rangle = \delta_{kl}$ and the fluctuations are colinear $\langle x_i^{(k)} x_j^{(l)}\rangle = r_k\delta_{kl}$, i.e., the part of their covariance matrix containing cross-correlations is diagonal. Then

$$\frac{\langle (x_j - f(x_i))^2 \rangle}{\langle x_j^2 \rangle} = \frac{1}{d}\sum_{k=1}^{d}(1 - r_k^2),$$

which simplifies in the case of identical correlation coefficients $r_k = r$ ($k = 1,\ldots,d$) to the *coefficient of nondetermination* for single variate variables, Eq. (2).

In the following discussion, and in accordance with common practice, we will use Eq. (1) also in the multivariate cases to define a *Pearson coefficient*, because of its similarity with the usual single-variate definition. However, in these cases several problems arise, which seemingly have not yet impeded widespread use.[8,11,12] First, the conditions colinearity and unit variance are generally not satisfied, thus invalidating the interpretation as a coefficient of nondetermination. This raises serious doubts regarding any conclusions drawn from this measure, particularly because any value for it can be obtained for a given ensemble by scaling single coordinates. Second, the Pearson coefficient is limited to detect *linear* correlations, i.e., it yields the coefficient of nondetermination regarding the best *linear* fit. *Nonlinear* fits, which can yield much lower coefficients of nondeterminations, are therefore not considered. This latter problem applies also to the one-dimensional case. Consider, e.g., two atoms oscillating in parallel direction, but with a 90° phase shift. They will give rise to a vanishing correlation matrix element $\langle \sin(\omega t)\ \sin(\omega t + \pi/2)\rangle = 0$, and, thus, this fully correlated motion would also not be detected. In configurational space, this motion generates an ensemble distributed along the perimeter of a circle, which cannot be captured by the Gaussian approximation implied in any formulation of correlated motion based on second moments.

Mutual Information

Among the measures of correlation between random variables, MI is singled out by its information theoretical background.[14] Accordingly, the joint probability distribution $p(\mathbf{x})$ is the product of the marginal distributions $p_i(\mathbf{x}_i)$,

$$p(\mathbf{x}) = \prod_{i=1}^{N} p_i(\mathbf{x}_i) \tag{3}$$

if and only if the components $\mathbf{x}_i$ are independent, i.e., uncorrelated. Because Eq. 3 can be rewritten as

$$\ln \frac{p(\mathbf{x})}{\prod_{i=1}^{N} p_i(\mathbf{x}_i)} = 0,$$

the ensemble-averaged deviation from the uncorrelated distribution is given by the MI,[14,15]

$$I[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N] = \int p(\mathbf{x}) \ln \frac{p(\mathbf{x})}{\prod_{i=1}^{N} p_i(\mathbf{x}_i)} d\mathbf{x}. \tag{4}$$

Only for fully uncorrelated motions, MI vanishes.

Evaluation of the right hand side of Eq. (4) relates MI to the more widely known measure of information content (entropy) $H[\mathbf{x}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$,

$$I[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N] = \sum_{i=1}^{N} H[\mathbf{x}_i] - H[\mathbf{x}]. \tag{5}$$

In contrast to the Pearson coefficient, this measure is scale-invariant. Even individual linear coordinate transformations in the d-dimensional subspaces, i.e. $(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N) \rightarrow (\mathbf{T}^{(1)}\mathbf{x}_1,\mathbf{T}^{(2)}\mathbf{x}_2,\ldots\mathbf{T}^{(N)}\mathbf{x}_N)$, as given by $d \times d$-matrices $\mathbf{T}^{(i)}$, leave the MI invariant, as little algebra shows. Here, we focus on the correlation between pairs of atoms,

$$I[\mathbf{x}_i,\mathbf{x}_j] = H[\mathbf{x}_i] + H[\mathbf{x}_j] - H[\mathbf{x}_i,\mathbf{x}_j]. \tag{6}$$

For higher order correlations we refer to Matsuda.[16]

Having established that the MI provides us with a well-defined and complete measure of correlation, we note that it yields values in the range $[0\ldots\infty)$, which is unfamiliar and has no obvious interpretation. Therefore, we will develop below an interpretation in terms of a coefficient of nondetermination, r_{MI}, which quantifies how well the best *nonlinear* model can describe the data. To this aim, we generalize the above one-dimensional linear case, for which the Pearson coefficient r directly allows this interpretation [Eq. (2)]. In particular, we suggest to relate $I[\mathbf{x}_i,\mathbf{x}_j]$ to a more intuitive Pearson-like coefficient $r_{\mathrm{MI}}[\mathbf{x}_i,\mathbf{x}_j]$ such that also in multidimensional and for nonlinear fit-functions f, the connection to the *coefficient of nondetermination* holds, i.e.,

$$1 - r_{\mathrm{MI}}[\mathbf{x}_i,\mathbf{x}_j]^2 = \frac{\langle (\mathbf{x}_j - f(\mathbf{x}_i))^2 \rangle}{\langle \mathbf{x}_j^2 \rangle}. \tag{7}$$

For fully correlated motions, this *generalized correlation coefficient* r_{MI} equals 1 and vanishes for fully uncorrelated motion.

To this end, we exploit that in the special case of Gaussian distributions ($d = 1$) or colinear Gaussian distributions of unit variance ($d = 3$) the Pearson correlation coefficient ($r = \langle\ \mathbf{x}_i \cdot \mathbf{x}_j\ \rangle = \langle |\mathbf{x}_i|\ |\ \mathbf{x}_j| \rangle$) captures all correlations. For this special case, one derives a one-to-one relationship between MI and the value of the Pearson correlation,

$$I_{\mathrm{Gauss}}[\mathbf{x}_i,\mathbf{x}_j] = -\frac{d}{2}\ln(1 - r^2). \tag{8}$$

Starting from this relationship, we define the *generalized correlation coefficient*, r_{MI}, as the Pearson coefficient of such a multidimensional Gaussian distribution, whose MI equals the one we wish to interpret. From Eq. (8),

$$r_{\mathrm{MI}}[\mathbf{x}_i,\mathbf{x}_j] = \{1 - \exp(-2I[\mathbf{x}_i,\mathbf{x}_j]/d)\}^{-1/2}, \tag{9}$$

which, as it is derived from the MI, contains all correlations. Therefore, for vectors of unit variance, $r_{\mathrm{MI}}[\mathbf{x}_i,\mathbf{x}_j]$ is always larger than $r[\mathbf{x}_i,\mathbf{x}_j]$. For multivariant cases, this rule may be violated because of the inconsistent scaling properties of r, which is repaired by r_{MI}. Note that the Gaussian distribution used to define r_{MI} will generally have a larger covariance than the original distribution, because Gaussians have the highest covariance compared with all possible distributions with the same MI.[17]

We now turn to numerically estimating the MI from a given ensemble or MD trajectory. For high-dimensional variables, crude approximations, such as cumulant expansions, are available.[17] For the correlation analysis of macromolecular dynamics, however, and in particular for the assessment of the correlated motion of atom pairs, density estimates for six-dimensional subspaces suffice. Approaches resting on k-nearest neighbor distances[18] or kernel density estimators[19] have proven to provide sufficiently accurate results for this purpose. The required accuracy is indeed very high, particularly for small correlations, for which the entropies involved nearly cancel out, hence small errors of the relatively large entropy terms lead to large errors in the estimated MI. This problem is aggravated because of the large slope of the transformation Eq. (9), in the low-correlation regime, which further amplifies errors. These strict accuracy requirements hold also for many other applications of the concept of MI, which recently instigated many developments.[18–24]

Linear Mutual Information (LMI)

The quite general and rigorous framework of MI also serves to single out nonlinear contributions to correlations. To this aim, recall that the Pearson coefficient suffers from two flaws: its inability to detect nonlinear correlations and its unwanted dependency on the relative orientation of the fluctuations. Thus, to separate the former from the latter, a reference quantity is required that suffers only from one of the two flaws. The LMI defined below serves this purpose. It has the additional advantage that its calculation does not require highly accurate and computationally demanding density estimates. Rather, it rests on a Gaussian approximation implied by the computationally much more efficient calculation of the covariance matrix, namely

$$g(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{(2\pi)^d \det(\mathbf{C}_{(ij)})} \exp\left(-\frac{1}{2}(\mathbf{x}_i,\mathbf{x}_j)\mathbf{C}_{(ij)}^{-1}(\mathbf{x}_i,\mathbf{x}_j)^{\mathrm{T}} \right),$$

with the pair-covariance matrix $\mathbf{C}_{(ij)} = \langle (\mathbf{x}_i,\mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i,\mathbf{x}_j)\rangle$. This Gaussian is the quasi-harmonic approximation to the canonical density of atomic motion. Thus, the MI, which can be computed analytically from this approximation, contains only linear correlations. The marginal probabilities are computed accordingly, using marginal-covariances $\mathbf{C}_{(i)} = \langle \mathbf{x}_i^T\mathbf{x}_i\rangle$. In contrast to the (general) MI, here, the required entropies are obtained analytically from the Gaussian density approximations, i.e., from the covariance matrices,

$$H(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{2}[2d(1 + \ln 2\pi) + \ln \det\mathbf{C}_{(ij)}].$$

Thus, from Eq. (5), the LMI,

$$I_{\mathrm{lin}}(\mathbf{x}_i,\mathbf{x}_j) = \frac{1}{2}[\ln \det\mathbf{C}_{(i)} + \ln \det\mathbf{C}_{(j)} - \ln \det\mathbf{C}_{(ij)}], \quad (10)$$

is obtained.

Similarly to the interpretation of (general) MI, the coefficient of nondetermination for the best multivariate *linear* fit is defined by Eqs. (7) and (9). LMI is a strict lower bound to MI, because the Gaussian distribution maximizes entropy under the constraints of a given mean and variance.[14] This is consistent with the definition of the generalized correlation coefficient and its interpretation, because the inclusion of nonlinear models will generally yield a higher coefficient of determination than restriction to linear models.

MD Simulations

Two MD trajectories were used. The first, termed GB1 and 200-ns long, was started from the crystal structure of the B1 domain of streptococcal protein G[25] [Protein Data Bank (PDB) entry 1PGB]. The protein was solvated in 4651 TIP4P water molecules using a cubic box. Four sodium ions were added to the simulation system to compensate for the net negative charge of the protein. The second trajectory, T4L, and 117-ns long, was started from the crystal structure of coliphage T4L M6I (PDB entry 150L chain D). The protein was solvated in 8898 TIP4P water molecules and 8 Cl^- counter ions using a rectangular box.

All MD simulations were conducted using the GROMACS simulation suite[26] and the OPLS all atom force field.[27] LINCS and SETTLE[28,29] were applied to constrain covalent bond lengths, allowing an integration step of 2 fs. Electrostatic interactions were calculated using the Particle-Mesh-Ewald method.[30,31] The temperature was kept constant by separately coupling (τ = 0.1 ps) the peptide and solvent to an external temperature bath.[32] The pressure was kept constant by weak isotropic coupling (τ = 0.1ps) to a pressure bath.[32]

Computation of Correlation Coefficients From MD Simulations

After 5-ns equilibration phase, coordinates were recorded every 10 ps. Thus, 19,500 and 11,200 coordinate sets were obtained and used for GB1 and T4L, respectively. Translational and rotational motions were removed by least squares fitting to the C_α-atoms of the respective crystal structures. The average structure $\langle\mathbf{r}\rangle$ was subtracted from the coordinates $\mathbf{r}$ to obtain centered atomic fluctuations $\mathbf{x}$. Correlations between fluctuations of the C_α-atoms were quantified by Pearson coefficients, Eq. (1), by linearized MI, Eq. (10), and by MI. For the latter, the density estimator by Kraskov et al.[18] was used with nearest neighbor parameter k = 6.

RESULTS AND DISCUSSION

Correlated Motion in Protein G

We first compare both correlation measures, the Pearson coefficient, Eq. (1), and the generalized correlation coefficient, r_{MI}, Eq. (9), for the B1 domain of protein G [Fig. 2(a)]. As expected, all correlations detected by the Pearson coefficient are also seen with the generalized correlation coefficient. Many additional correlations are revealed by r_{MI}, however, that are not revealed by the Pearson coeffi-

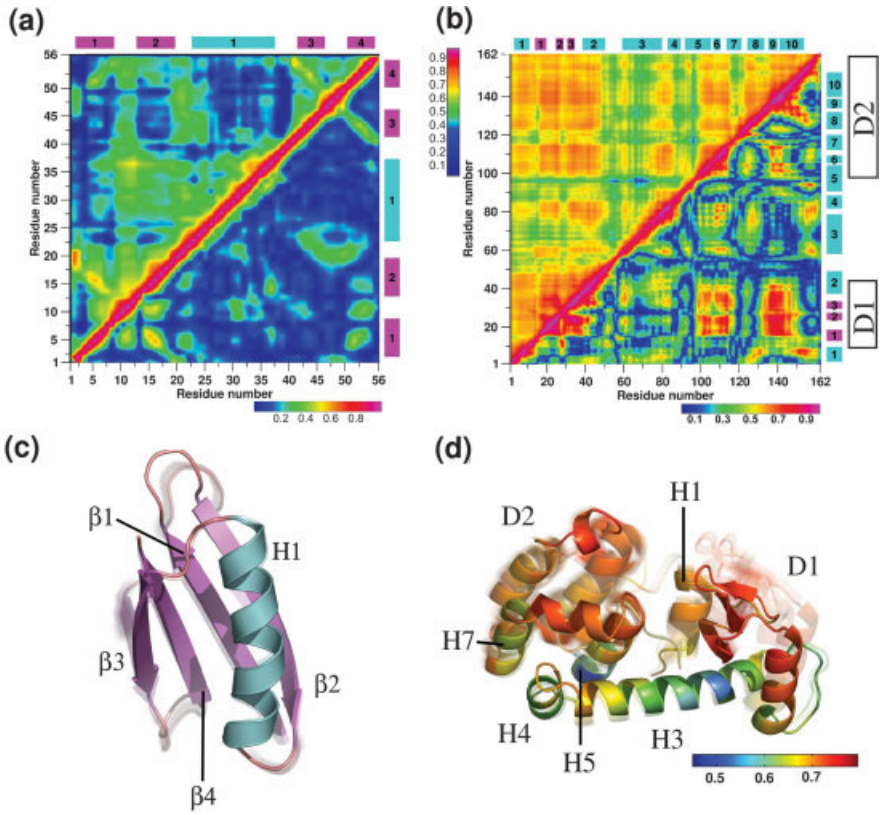

Fig. 2. **a,b:** Generalized correlation coefficient r_{MI} (upper left triangle) and Pearson coefficient $|r|$ (lower right triangle) correlation matrices for (a) the B1 domain of protein G (GB1) and for (b) T4 lysozyme (T4L). The strength of the computed correlation between two respective residues is color-coded (see color bars); note that different color mappings are used to enhance contrast. Secondary structure elements are indicated by bars in magenta (β-sheets) and cyan (α-helices). **c,d:** Structure and superimposed three frames from the GB1 and lysozyme (T4L) trajectory, respectively, indicating the amplitude of the observed motion. **d:** For every residue, the mean correlation with residues of the two domains D1 (15–46) and D2 (100–160) was computed and color-coded.

cient. Furthermore, as will be analyzed in detail below, the purely geometrical (orientational) perturbation of the Pearson coefficient creates patterns in the Pearson matrix that actually are unrelated to any correlation and in this sense artificial.

Correlations detected by both methods are found along the diagonal and in two bands perpendicular to the main diagonal. The latter are attributed to the hydrogen bonded contacts between different strands of the four-stranded β-sheet. The correlations between strands β_1-γ_2 and β_3-β_4 are pronounced, whereas the correlations between hydrogen bonding partners of the central neighbors β_1-β_4, showing up as band parallel to the diagonal, are weaker.

The broad region of high correlation along the main diagonal between residues 22 and 38 is caused by the close packing of residues in the α-helix. The correlation between hydrogen bonded residues in the helix is slightly weaker than correlation between opposing C_α-atoms in β-sheets. The reason for this is that in β-sheets, both neighbors of the C_α-atom are tightly hydrogen-bond coupled to one residue of the parallel strand, whereas in the helix the two neighbors couple to two different residues in opposite direction.

New, so far undetected correlations, are seen in the generalized correlation matrix. These include—less pronounced, but significant—correlations between the α-helix and the first double strand of the β-sheet (β_1, β_2), which are absent for the second double strand (β_3, β_4). This finding can also be explained in terms of geometrical proximity. The helix of GB1 traverses diagonally one half of the β-sheet; starting above residues 50 and 1 of strand β_4 and β_1, respectively, it extends outward ending near residue 13 of β_2 [cf. Fig. 2(c)]. Therefore, the larger part of the helix is located far from strands (β_3, β_4) and closely to strands (β_1, β_2), yielding correlations with the latter only, whereas the residues in the preceding loop and the adjacent part of the helix are close enough to (β_3, β_4) to also cause correlations with these strands.

In summary, the largest correlated motions observed in GB1 are rather caused by geometrical proximity than by collective conformational motion, and are, in this sense, trivial. These large correlations are, not surprisingly, captured by both measures, Pearson coefficient and MI. However, whereas the Pearson coefficient focuses on the correlation inside secondary structure elements, MI re-

veals many new and nontrivial medium strong correlations between different secondary structure elements.

Correlated Motion in T4 Lysozyme

The single protein domain GB1 characterized above is intrinsically rigid. Now we turn to T4L, which exhibits two well-separated domains and significant conformational interdomain motions.[33,34] Experimental and theoretical studies have shown that these domain motions are essential for the function of this enzyme, allowing the substrate to enter and the products to leave the active site.[35–38] Atomic correlations have been analyzed extensively for lysozyme using the Pearson coefficient matrix [Fig. 4(b), lower right).[11] Here, we have calculated the MI-based generalized correlation coefficient matrix and focus on the new features this analysis has revealed.

Figure 2(b) (upper triangle) shows that the MI successfully quantifies the highly correlated motion within and between the two domains, D1, residues 13–50, and D2, residues 100–162, of T4L (cf. Fig. 2). The second domain (D2) moves as two rigid blocks, formed by H4–H6 and H8–H9, respectively, which are weakly linked by residues 118–121. Interestingly, the less correlated linker residues are part of H7 and not, as one might expect, part of the loop region between helices.

Furthermore, the generalized correlation matrix shows that the interdomain motion is not just a simple hinge motion[36,37,39] with H3 and H4 (residues 62–90) forming the hinge region, as one might expect. In fact, a typical hinge motion would imply smaller correlations for the hinge, as indeed found for residues 85–98 (part of H4 and H5) and residues 62–74 (part of H3). Instead, part of the hinge regions, namely adjacent parts of H3 and H4 (residues 75–84) correlate strongly with the overall domain motion, which would not be the case for a simple hinge motion.

The N-terminal helix H1, which contains active site residues, moves correlated with both domains D1 and D2. However, in contrast to these domains, it shows only weak correlation to the aforementioned linker region around H4. These results are consistent with a previously conducted principal component analysis (PCA),[40] where the conformational motion of T4L could be described by a rigid body closure and twist motion of domains D1 and D2. That study showed rigid co-motion of H1 and D2 for the closure motion, and, for the twist motion, H1 moves with D1. This splitting up of the H1 correlation can now be understood by considering the nonlinear contributions to the overall correlation obtained as difference between MI and LMI (Fig. 3). As can be seen, the correlations between domains D1 and D2 are mostly linear in nature, whereas the correlation of H1 with both domains has significant nonlinear contributions. This explains why the rather nonlinear correlation of H1 with the domains was found to be distributed over two *linear* principal modes.[40]

In contrast to the complete quantification of correlated motions by the generalized correlation coefficient, the Pearson coefficient picks up only parts of these correlations, and many remain undetected. This can give rise to a

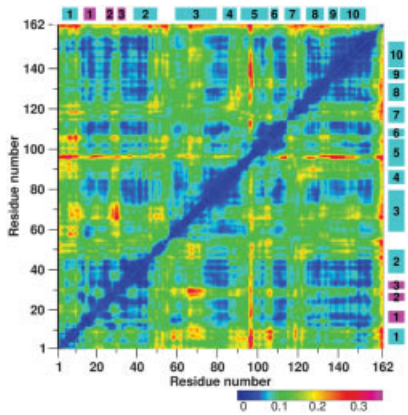

Fig. 3. Matrix of purely nonlinear contributions $r_{MI}-r_{LMI}$ to the correlations between atom pairs in T4L; see text.

rather inconsistent picture, i.e., patterns in the results not reflecting patterns of correlation, which is particular pronounced for T4L [cf. Fig. 2(b), lower triangle]. Although the two domains move as relatively rigid units, the Pearson coefficient quantifies the correlations within the domains rather incompletely. Whereas the Pearson coefficient does show correlations within the first part of D1, the correlations between the first part of D1 and its last 10 residues (40–50) seen by the generalized correlation coefficient are missing. Moreover, most correlations within domain D2 are undetected. A particularly striking and obvious inconsistency would be the violation of transitivity, i.e., two regions between which no correlated motions are detected, but which are both correlated to a third one. Such a situation is indeed purported by the Pearson correlation coefficient, which indicates a high correlation of D1 with the two regions, residues 100–118 and residues 130–150 of D2, but misleadingly low correlation between these two regions. Finally, the Pearson coefficient does not detect H1 to be correlated with D1, and detects only a small fraction of the correlations between H1 and residues of D2. In some instances, the Pearson correlation measure yields higher values than the generalized correlation, which should, in principle, not happen. The two possible reasons, the scaling dependency of the Pearson measure or numerical inaccuracies in the estimation of MI, are discussed further below.

Thus, the proposed generalized correlation coefficient based on MI yields a much more complete picture of the correlated motions which is consistent with—and extends—previously applied PCA.[40] In particular, whereas the Pearson correlation coefficient captures most of the correlated motions within the B1 domain of protein G, it misses many pronounced correlated motions of lysozyme, which involve all active site residues and are likely to be functionally important. The nature of this failure and the question under which conditions it is to be expected, deserves closer inspection.

Analysis of the Failures of the Pearson Coefficient

Figure 4 compares as scatter plots all elements of the generalized correlation coefficient matrices with the respec-

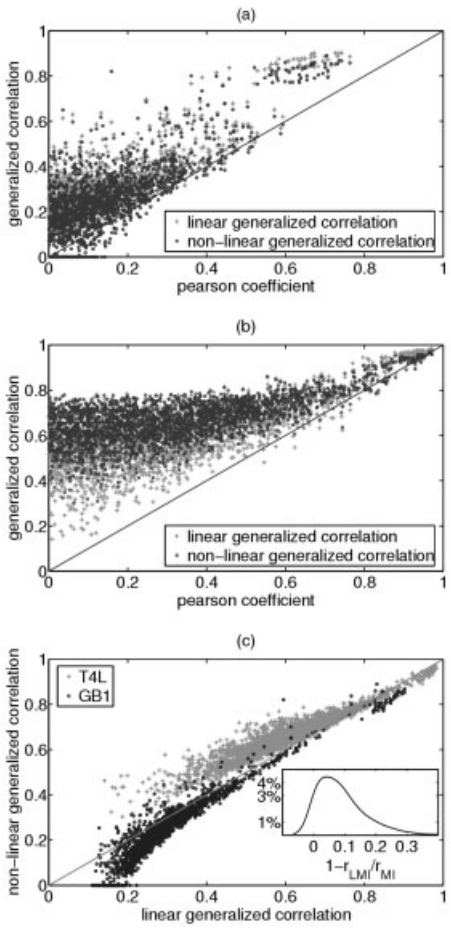

Fig. 4. Comparison of MI-based correlation measures with Pearson correlation coefficients. For pairs of C_α–atoms of (**a**) GB1 and (**b**) T4L, both generalized correlation coefficients r_{MI} (black circles) and r_{LMI} (gray crosses) are plotted against the Pearson correlation coefficient. **c:** Comparison between linear and nonlinear MI. For GB1 (black) and T4L (gray), the generalized correlation coefficients computed from LMI are plotted against nonlinear generalized correlation. For T4L, the inset shows a histogram of the differences between both coefficients, with maximum of the distribution at 0.04 and a mean of 0.09.

tive elements of the Pearson correlation matrices, both shown in Figure 2(a, b). Results are shown for both GB1 and T4L [Fig. 4(a, b)]. For large correlations ($r_{\mathrm{MI}} \geq 0.8$), the Pearson coefficient r and the generalized correlation coefficient r_{MI} give comparable results. For less correlated motions ($r_{\mathrm{MI}} \leq 0.8$), the Pearson coefficient rarely captures the full correlation, and often underestimates r_{MI} considerably, yielding any value between zero and r_{MI}. In fact, as quantified by the average underestimation $[\Sigma_{ij}|r_{ij}|/r^{ij}_{\mathrm{MI}}]/N = 0.48$, only less than half of the correlations are revealed by the Pearson coefficient. Below, we analyze the causes for the erratic occurrences of their drastic underestimation.

As discussed in Materials and Methods, possible causes are a) the dependence on the relative orientation of the fluctuations, b) the presence of nonlinear correlations, and c) lack of scaling invariance. We will demonstrate below that the dependence on direction is in fact the main cause of the underestimation enhanced by the presence of nonlinear correlations.

We start the analysis by separating the effect of nonlinear correlations from the purely linear contributions. To this end, Figure 4 compares the generalized correlation coefficient discussed above with the corresponding coefficient based on *linear* MI (see Materials and Methods). As can be seen, both agree well for GB1 except for numerical inaccuracies within the low-correlation regime. In contrast, clear deviations for lysozyme point toward significant nonlinear correlations. Indeed, as qualified by the histogram of deviation (inset) or quantified by $\Sigma_{ij}(r_{\mathrm{MI}} - r_{\mathrm{LMI}})/r^{ij}_{\mathrm{MI}}]/N = 0.09$, the nonlinear part of the correlation contributes up to 10% to the overall correlation and, therefore, accounts for a significant part of the correlation not described by the Pearson coefficient (cf. crosses in Fig. 4). Because both r_{LMI} and r rely on the linear quasi-harmonic approximation, the remaining ca. 40% of the undetected correlations—in fact the largest part—cannot be explained by nonlinear effects.

To quantify the (geometrical) effect of relative orientation of the atomic fluctuations on the Pearson coefficient, the latter was separated into correlations of distances,

$$r_{\mathrm{abs}}[\mathbf{x}_i,\mathbf{x}_j] = \langle |\mathbf{x}_i|\,|\mathbf{x}_j| \rangle / (\langle \mathbf{x}_i \rangle \langle \mathbf{x}_j \rangle^2)^{1/2}, \tag{11}$$

and average colinearity

$$r_{\mathrm{dir}}[\mathbf{x}_i,\mathbf{x}_j] = \left\langle \left| \frac{\mathbf{x}_i}{|\mathbf{x}_i|} \cdot \frac{\mathbf{x}_j}{|\mathbf{x}_j|} \right| \right\rangle. \tag{12}$$

Figure 5 compares the correlations of distances, r_{abs}, with both the linear generalized correlation coefficient r_{LMI} (black) as well as the Pearson coefficient r (red). As can be seen, r_{abs} is more closely linked to r_{LMI} than to the Pearson coefficient, as is also quantified by correlation coefficients of 0.88 versus 0.64, respectively. In contrast, the average colinearity is more linked to the Pearson coefficient (correlation coefficients 0.47 versus 0.87, respectively; data not shown), thus confirming that the relative orientation of the atomic fluctuations perturbs the Pearson coefficient considerably. Indeed, as shown in Figure 6, the average orientation is closely linked to the divergence of the Pearson coefficient from the generalized correlation, quantified by a correlation coefficient of −0.78. Knowledge of the relative orientations of the fluctuations alone, therefore, allows prediction of when the Pearson coefficient will fail to detect correlations. For high colinearity, the Pearson coefficient quantifies the correlation relatively well, whereas it systematically underestimates the correlation in cases where the fluctuations are nearly perpendicular to each other.

Interestingly, additional consideration of the generalized correlation coefficient in Figure 6 (color-coded) shows that the very high correlation in lysozyme coincides exclusively with colinear motions—in which case the Pearson coefficient performs quite well. This is explained by the fact that, for the case of protein dynamics, these high correlations can only arise from atoms confined within secondary structure elements. That only medium strong

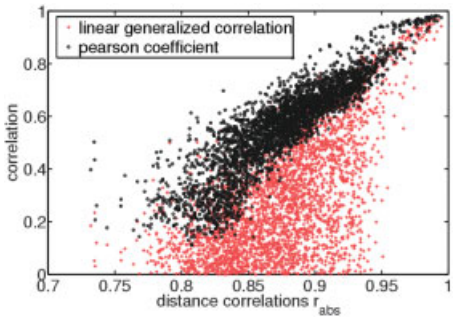

Fig. 5. Distance correlations r_{abs} compared with the two linear correlation measures r_{LMI} (black) and Pearson coefficient (red).

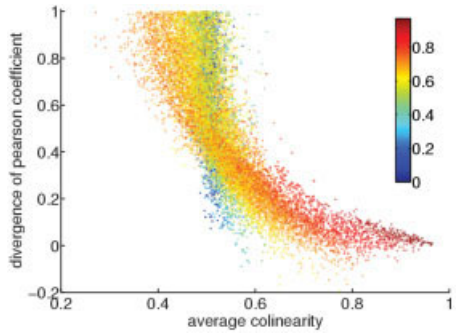

Fig. 6. Relative (linear) divergence of the Pearson coefficient $\Delta|r| = 1 - |r|/r_{LMI}$ as a function of the average colinearity r_{dir}. The colors quantify full (nonlinear) correlation r_{MI}.

correlations are missed by the Pearson coefficient is, therefore, rather attributable to the specific properties of protein dynamics and not to a merit of the Pearson coefficient. Similarly, the high performance of the Pearson coefficient at high correlations may hold only for dynamics and is not a general property of the Pearson coefficient.

The only defect in the definition of the Pearson coefficient not discussed so far is its lack of scaling invariance. Closer inspection of Figure 4 reveals atom pairs for which the Pearson coefficient is slightly higher than the r_{LMI}. Because both measures are based on the same linear, i.e., harmonic approximation, this cannot be explained with numerical estimation inaccuracies. We suggest as the likely cause the improper scaling behavior of the Pearson coefficient, which overestimates the correlation. However, because the effect is small, we were not able to separate it from the other effects discussed above, so that this hypothesis could not be proven.

We finally discuss the numerical inaccuracies mentioned above and described in Materials and Methods. For certain atom pairs, the nonlinear correlation is actually lower than the linear correlation (Fig. 4), which should not occur because LMI is a strict lower bound to nonlinear MI (see Materials and Methods). However, here the MI is estimated from a finite number of frames, which implies statistical inaccuracies. Because MI is a difference of relatively large entropies, the relative error increases for small correlations, which explains the deviations seen in Figure 4 for low MI $r_{LMI} < 0.3$. At higher correlations ($r_{MI} \gtrsim 0.7$), a small systematic underestimation of MI is observed, as discussed in Kraskov et al.[18] Taken together, accuracy can be enhanced by using the larger value of both the (analytical) linear and the (numerical) nonlinear MI.

CONCLUSIONS

We have derived a generalized correlation measure based on MI, which allows for complete characterization and quantification of atomic correlations in proteins and other macromolecular motion. It provides a consistent framework for analyzing correlations between coordinates, atoms, and groups of atoms, and thereby overcomes the problems of the usually employed Pearson correlation coefficient.

First, both linear and nonlinear contributions to correlation are accounted for. Moreover, a linearized generalized correlation coefficient was derived within the framework of MI which allowed separation of linear and nonlinear contributions to correlation. For T4L, the latter account for approximately 10% of all correlations. Second, our generalized correlation coefficient does not suffer from the artifacts of the established method that originate from the relative orientation of the atomic fluctuations. This purely geometrical artifact of the Pearson coefficient typically leads to underestimation of the correlations by more than 40%. Taken together, more than 50% of the correlations remain undetected by the established method, but are fully accounted for by the generalized correlations coefficient.

Application to two proteins, the B1 domain of protein G and coliphage T4L, revealed new information on their functionally relevant collective dynamics. In particular for lysozyme, the established characterization of the domain motion in terms of a hinge motion has been extended toward a more complex pattern of collective motions. This pattern is not revealed by the conventional Pearson coefficient matrix, which, in addition, conveyed misleading information.

The enhanced characterization of the collective motion provided by the generalized correlation matrix also complements the analysis of collective motions with PCA. For example, the assessment of nonlinear correlations presented here can explain the previous finding by PCA that for T4L the helix H1 moves either rigidly together with domain D1, as shown by the first principal component, or, for the second principle component, H1 moves together with domain D2.[40]

Overall, particularly many interdomain motions were revealed by the generalized correlation coefficient. In contrast, the Pearson correlation coefficient turned out to focus at the local correlations, which often are attributable to spatial proximity within secondary structure elements and in this sense virtually trivial. Particularly the interdomain motions, however, tend to exhibit nonlinear correlations, which can now be captured by the generalized correlation.

We note that the presented definition of MI can be generalized to higher dimensions. Accordingly, correlations between groups of atoms can also be quantified, e.g., between a ligand and selected residues of its binding

pocket. To this aim, the application of linearized MI is straightforward. For the nonlinear MI, the numerical estimation used here may become inaccurate for larger numbers of atoms per group. In this case, parametric entropy estimators will be superior.[14]

The generalized correlation coefficient developed in this work is widely applicable to the exponentially growing amount of configurational ensembles provided by MD simulations and from other sources such as NMR or CONCOORD.[41] This method will thus allow for the detection and characterization of a large number of new functionally important protein motions. Moreover, it facilitates direct comparison with experimental data, e.g., from X-ray diffusive scattering, NMR, or, via entropy, from calorimetry. The method has been implemented within the GROMACS simulation suite[26] and can be downloaded from http://www.mpibpc.mpg.de/groups/grubmueller/olange/gencorr.html.

ACKNOWLEDGMENTS

The authors thank Bert L. de Groot for providing the T4 lysozyme trajectory and for helpful discussions; and Frauke Gräter for carefully reading the manuscript.

REFERENCES

1. Agarwal PK, Billeter SR, Rajagopalan PTR, Benkovic SJ, Hammes-Schiffer S. Network of coupled promoting motions in enzyme catalysis. Proc Natl Acad Sci USA 2002;99:2794–2799.
2. Scheer A, Cotecchia S. Constitutively active G protein-coupled receptors: potential mechanisms of receptor activation. J Recept Signal Transduct Res 1997;17:57–73.
3. Cross RL. Our primary source of ATP. Nature 1994;370:594–595.
4. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.
5. Forman-Kay JD. The 'dynamics' in the thermodynamics of binding. Nat Struct Biol 1999;6:1086–1087.
6. Lee AL, Wand AJ. Microscopic origins of entropy, heat capacity and the glass transition in proteins. Nature 2001;411:501–504.
7. Case DA. Molecular dynamics and NMR spin relaxation in proteins. Acc Chem Res 2002;35:325–331.
8. Meinhold L, Smith JC. Fluctuations and correlations in crystalline protein dynamics: a simulation analysis of staphylococcal nuclease. Biophys J 2005;88:2554–2563.
9. Mayer KL, Earley MR, Gupta S, Pichumani K, Regan L, Stone MJ. Covariation of backbone motion throughout a small protein domain. Nat Struct Biol 2003;10:962–965.
10. Lange OF, Grubmüller H, de Groot BL. Molecular dynamics simulations of protein G challenge NMR-derived correlated backbone motions. Angew Chem Int Ed 2005;44:3394–3399.
11. Hünenberger PH, Mark AE, van Gunsteren WF. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. J Mol Biol 1995;252:492–503.
12. Ichiye T, Karplus M. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular-dynamics and normal mode simulations. Proteins 1991;11:205–217.
13. Briki F, Genest D. Canonical-analysis of correlated atomic motions in DNA from molecular-dynamics simulation. Biophys Chem 1994;52:35–43.
14. Cover TM, Thomas JA. Elements of information theory. New York: John Wiley & Sons; 1991.
15. Fraser AM, Swinney HL. Independent coordinates for strange attractors from mutual information. Phys Rev A 1986;33:1134–1140.
16. Matsuda H. Physical nature of higher-order mutual information: intrinsic correlations and frustration. Phys Rev E 2000;62:3096–3102.
17. Hyvarinen A, Oja E, Karhunen J. Independent component analysis. New York: Wiley; 2001.
18. Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information. Phys Rev E 2004;69:066138.
19. Shwartz S, Zibulevsky M, Schechner YY. Fast kernel entropy estimation and optimization. Signal Process 2005;85:1045–1058.
20. Grassberger P. Finite-sample corrections to entropy and dimension estimates. Phys Lett A 1988;128:369–373.
21. Moon YI, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators. Phys Rev E 1995;52:2318–2321.
22. Roulston MS. Estimating the errors on measured entropy and mutual information. Physica D 1999;125:285–294.
23. Darbellay GA, Vajda I. Entropy expressions for multivariate continuous distributions. IEEE Trans Inf Theory 2000;46:709–712.
24. Learned-Miller EG, Fisher JW. ICA using spacings estimates of entropy. J Mach Learn Res 2004;4:1271–1295.
25. Gallagher T, Alexander P, Bryan P, Gilliland GL. Two crystal-structures of the B1 immunoglobulin-binding domain of streptococcal protein-G and comparison with NMR. Biochemistry 1994;33:4721–4729.
26. Lindahl E, Hess B, Van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 2001;7:306–317; http://www.gromacs.org.
27. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
28. Miyamoto S, Kollman PA. SETTLE: an analytical version of the SHAKE and RATTLE algorithms for rigid water models. J Comp Chem 1992;13:952–962.
29. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. J Comp Chem 1997;18:1463–1472.
30. Darden T, York D, Pedersen L. Particle mesh Ewald: an $N \cdot \log(N)$ method for Ewald sums in large systems. J Chem Phys 1993;98:10089–10092.
31. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. J Chem Phys 1995;103:8577–8593.
32. Berendsen HJC, Postma JPM, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys 1984;81:3684–3690; a typical GROMACS cite.
33. Matthews BW, Remington SJ. The three-dimensional structure of lysozyme from bacteriophage T4. PNAS 1974;71:4178–4182.
34. McCammon JA, Gelin B, Karplus M, Wolynes PG. The hinge bending mode in lysozyme. Nature 1976;262:325–326.
35. Kuroki R, Weaver LH, Mathhews BW. A covalent enzyme-substrate intermediate with saccharide distortion in a mutant T4 lysozyme. Science 1993;262:2030.
36. Faber HR, Matthews BW. A mutant T4 lysozyme displays five different crystal conformations. Nature 1990;348:263–266.
37. Mchaourab HS, Oh KJ, Fang CJ, Hubbell WL. Conformation of T4 lysozyme in solution. Hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. Biochemistry 1997;36:307–316.
38. Lu HP. Single-molecule spectroscopy studies of conformational change dynamics in enzymatic reactions. Curr Pharm Biotechnol 2004;5:261–269.
39. Hayward S, Berendsen HJC. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. Proteins 1998;30:144–154.
40. de Groot BL, Hayward S, van Aalten DMF, Amadei A, Berendsen HJC. Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. Proteins 1998;31:116–127.
41. de Groot BL, van Aalten DMF, Scheek RM, Amadei A, Vriend G, Berendsen HJC. Prediction of protein conformational freedom from distance constraints. Proteins 1997;29:240–251.