Anharmonicity, mode-coupling and entropy in a fluctuating native protein

A. Kabakçıoğlu[†], D. Yüret[*], M. Gür[*], B. Erman[*,†]
Colleges of Sciences[†] and Engineering[*], Koç University, Sarıyer, 34450, İstanbul, Turkey

We develop a general framework for the analysis of residue fluctuations that simultaneously incorporates anharmonicity and mode-coupling in a unified formalism. We show that both deviations from the Gaussian model are important for modeling the multidimensional energy landscape of the protein Crambin (1EJG) in the vicinity of its native state. The effect of anharmonicity and mode-coupling on the fluctuational entropy is on the order of a few percent.

PACS numbers: 87.14.E-, 87.15.-V, 87.85.J-

Residue fluctuations of a protein around its native state reveal information that bridges the molecule's structural and functional properties. At the lowest order, these fluctuations can be treated as a collection of independent harmonic modes, yielding "Elastic Network Models" [1, 2]. On the other hand, it is known that the slowest oscillatory modes of a protein are strongly anharmonic [3–5], in contrast with the assumption underlying ENMs. The coupling between different modes is another aspect of protein dynamics that is believed to be relevant for the information/energy transfer between different parts of the molecule [6, 7] and not captured by the harmonicity assumption.

Sampling the time evolution of a protein by using molecular dynamics reveals a multivariate probability distribution function (pdf) $f(\boldsymbol{\Delta}\mathbf{R})$ for the deviations of atoms (assume there are N of them) from equilibrium coordinates, i.e. $\Delta R_i = R_i - R_i^{eq}$, $i = 1, \ldots, 3N$. We here adopt a coarse-grained representation of this pdf where only C^α atoms are considered, so that N is also the number of residues. Accordingly, R_i^{eq} are the mean C^α coordinates corresponding to the average configuration of the protein during the part of the trajectory that is used for the calculations. Since the deviations from this free energy minimum should be harmonic for sufficiently small amplitudes, Hermite polynomials - which are orthogonal *wrt* a Gaussian weight function - constitute a natural basis for representing $f(\boldsymbol{\Delta}\mathbf{R})$. First, following Ref. [3, 8], we perform the transformation

$$\boldsymbol{\Delta}\mathbf{r} \;=\; \langle \boldsymbol{\Delta}\mathbf{R}\boldsymbol{\Delta}\mathbf{R}^T \rangle^{-1/2} \boldsymbol{\Delta}\mathbf{R}\ . \tag{1}$$

This diagonalizes the covariance matrix $\Gamma \equiv \langle \boldsymbol{\Delta}\mathbf{R}\boldsymbol{\Delta}\mathbf{R}^T \rangle$ ($\langle\cdot\rangle$ represents averaging over the trajectory) and would give the normal modes of the protein if fluctuations were harmonic. Otherwise, the distribution function for $\{\Delta r_i\}$ in its most general form, can be expressed as [9]

$$f(\boldsymbol{\Delta}\mathbf{r}) = \frac{1}{\sqrt{(2\pi)^{3N}}} e^{-\frac{1}{2}\sum_{i=1}^{3N} \Delta r_i^2} \left[1 + \sum_{\nu=3}^{\infty} \mathbf{C}_\nu \cdot \mathbf{H}_\nu(\boldsymbol{\Delta}\mathbf{r}) \right] \tag{2}$$

where $\mathbf{C}_\nu$ (constant) and $\mathbf{H}_\nu$ (derived below) are tensors of rank ν, and the dot product refers to $\sum_{ij..k} C_\nu^{ij..k} H_\nu^{ij..k}$. The fluctuations $\{\Delta r_i\}$ in this mode space spanned by the eigenvectors of Γ are meanless, i.e., $\langle \Delta r_i \rangle = 0$, and decoupled at the lowest (second) order, i.e.,

$$\langle \Delta r_i^T \Delta r_j \rangle = \delta_{ij}\ . \tag{3}$$

A purely harmonic model is given by $\mathbf{C}_\nu = 0$, $\forall \nu$. Note that, the atomic fluctuations corresponding to a given mode can easily be calculated by setting to zero all the eigenvalues except the one of interest, followed by a back transformation of Eq. (1).

Tensor Hermite polynomials can be obtained by successive differentiation using Rodrigues' formula:

$$H_\nu^{ij..k}(\boldsymbol{\Delta}\mathbf{r}) \;=\; \frac{(-1)^\nu}{g(\boldsymbol{\Delta}\mathbf{r})} \nabla^{ij..k} g(\boldsymbol{\Delta}\mathbf{r})\ . \tag{4}$$

Above, $g(\mathbf{x}) = (2\pi)^{3N/2} \exp(-\mathbf{x}^2/2)$ is the multidimensional Gaussian distribution and $\nabla^{ij..k} = \nabla^i \nabla^j \,..\, \nabla^k$ is the gradient tensor with $\nabla^i \equiv \partial/\partial x_i$. The tensor coefficients that appear in $f(\boldsymbol{\Delta}\mathbf{r})$ follow from the orthogonality relation as

$$\mathbf{C}_\nu \;=\; \frac{1}{\nu!} \int_{-\infty}^{\infty} \mathbf{H}_\nu(\mathbf{x}) f(\boldsymbol{\Delta}\mathbf{r})\, \mathbf{d}\boldsymbol{\Delta}\mathbf{r} = \langle \mathbf{H}_\nu(\boldsymbol{\Delta}\mathbf{r}) \rangle / \nu! \tag{5}$$

Therefore, the problem reduces to obtaining the expectation values of the polynomial tensor elements for the system. At the lowest nonvanishing order they read

$$\begin{aligned} \mathbf{H}_3^{111}(\mathbf{x}) \;&=\; x_1^3 - 3x_1 \\ \mathbf{H}_3^{112}(\mathbf{x}) \;&=\; x_1^2 x_2 - x_2 = \mathbf{H}_3^{121}(\mathbf{x}) = \mathbf{H}_3^{211}(\mathbf{x}) \\ \mathbf{H}_3^{123}(\mathbf{x}) \;&=\; x_1 x_2 x_3 = \mathbf{H}_3^{213}(\mathbf{x}) = \cdots \end{aligned} \tag{6}$$

Higher order tensor elements can be calculated using a diagrammatic technique. A graphical representation of H_4 in one and two dimensions is given in Fig.1.

The inclusion of mode-coupling necessitates consideration of mixed indices (nondiagonal tensor elements). Here, we focus on the coupling between mode pairs and ignore threesome and higher order mixing, i.e., we consider only the bi-polynomials $\mathbf{H}_\nu^{i_1 i_2 \ldots i_\nu}(\Delta r_k, \Delta r_l)$ with $i_m \in \{k, l\}$, $k, l = 1, 2, \ldots, 3N$. At first sight, estimating the contribution of mode-coupling even at this lowest level appears to be a formidable task, because the number of distinct expectation values to be extracted from

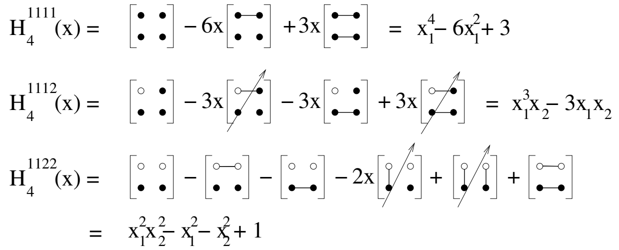
H4 1111(x) = − 6x + 3x = x1^4 − 6x1^2 + 3
H4 1112(x) = − 3x − 3x + 3x = x1^3 x2 − 3x1 x2
H4 1122(x) = − − − 2x + +
= x1^2 x2^2 − x1^2 − x2^2 + 1

FIG. 1: Graphical representation of $\mathbf{H}_4(\mathbf{x})$ tensor elements in two dimensions. Terms that vanish by virtue of Eq. (3) are crossed.

the data grows combinatorially. We show below that, the factorization property of the off-diagonal tensor elements and the orthogonality of the modes at the second order bring a significant reduction in complexity, which we exploit to investigate the impact of anharmonicity and mode-coupling separately on the protein dynamics.

Mode-coupling corrections

The value of a tensor element $\mathbf{H}_\nu^{i_1 i_2 \ldots i_\nu}(\mathbf{\Delta r})$ does not depend on the order of the indices due to the commutativity of the gradient operator, $\nabla_k \nabla_l - \nabla_l \nabla_k = 0$. Therefore,

$$\mathbf{H}_\nu^{i_1 i_2 \ldots i_\nu}(\mathbf{\Delta r}) \; = \; \mathbf{H}_\nu^{p}(\Delta r_k, \Delta r_l)$$

where p is the number of indices equal to k (and the remaining $\nu - p$ indices are equal to l). The fact that the covariance matrix in the normal basis is diagonal further implies that

$$\mathbf{H}_\nu^{p}(\mathbf{\Delta r}) \; = \; H_p(\Delta r_1) \times H_{\nu-p}(\Delta r_2) \qquad (7)$$

as is also evident from the Rodrigues's formula in Eq. (4).

Combining Eq. (5) and Eq. (7), the Hermite expansion in Eq. (2) can be cast into the following form:

$$\begin{aligned} f(\mathbf{\Delta r}) \; &= \; \frac{1}{\sqrt{(2\pi)^N}} e^{-\sum_i \Delta r_i^2/2} \Bigg[1 + \sum_i \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \Big\langle H_\nu(\Delta r_i) \Big\rangle H_\nu(\Delta r_i) \\ &+ \; \sum_{i \neq j} \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \sum_{p=1}^{\nu-1} \binom{\nu}{p} \Big\langle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \Big\rangle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) + \sum_{i \neq j \neq k} \cdots \Bigg] \end{aligned} \qquad (8)$$

The first term in Eq. (8) corresponds to a purely harmonic model given by the Gaussian probability distribution

$$f_0(\mathbf{\Delta r}) \; = \; \frac{1}{\sqrt{(2\pi)^N}} e^{-\sum_i \Delta r_i^2/2} \; . \qquad (9)$$

This is the starting point for most of the past studies on protein fluctuations [10]. The next term in Eq. (8) is appreciable when the fluctuations are anharmonic, but gives no information about mode-coupling. In fact, the most general mode-amplitude distribution of an anharmonic model composed of decoupled modes is

$$\begin{aligned} f_1(\mathbf{\Delta r}) \; &= \; \frac{1}{\sqrt{(2\pi)^N}} \; e^{-\sum_i \Delta r_i^2/2} \times \\ & \prod_i \left[1 + \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \Big\langle H_\nu(\Delta r_i) \Big\rangle H_\nu(\Delta r_i) \right] (10) \end{aligned}$$

The approximation to the true distribution given in Eq. (10) is named f_1 in order to remind the reader that it qualitatively improves on the Gaussian approximation f_0 of Eq. (9). The difference between the full pdf given in Eq. (8) and the approximation f_1 is the mode-coupling corrections such as

$$\langle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \rangle - \langle H_p(\Delta r_i) \rangle \langle H_{\nu-p}(\Delta r_j) \rangle \; \neq \; 0$$

and higher order cumulants. Note that, marginal distributions are transparent to such corrections

$$f(\Delta r_i) \; \equiv \; \int_0^\infty \prod_{j \neq i} d\Delta r_j \; f(\mathbf{\Delta r}) \qquad (11)$$

as a merit of the orthogonality relation in Eq. (5). Therefore, even if the marginal distributions are reproduced to good accuracy, the multidimensional free-energy landscape of the protein may still be very different from that implied by a model based on Eq. (10). We demonstrate below that this is the case for the protein Crambin (1EJG). To this end, we improve the approximation in Eq. (10) one step further and approximate $f(\mathbf{\Delta r})$ by

$$f_2(\mathbf{\Delta r}) \; \equiv \; f(\mathbf{\Delta r}) - \sum_{i \neq j \neq k} \cdots \; , \qquad (12)$$

i.e., the part of the Hermite expansion spelled out in Eq. (8) which takes into account the mode-coupling corrections at the lowest order they appear while ignoring cubic and higher-order terms.

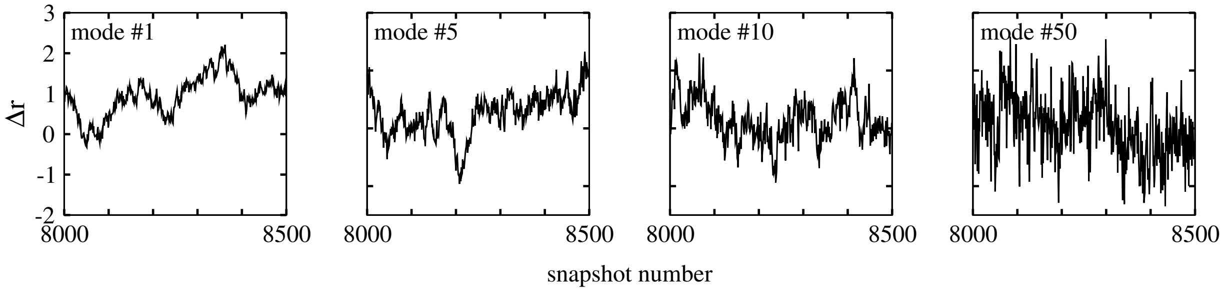

FIG. 2: Time plots of the slowest 1^{st}, 5^{th}, 10^{th}, and 50^{th} modes between timesteps 8000-8500.

Crambin molecular dynamics: a test ground

Crambin (Protein Data bank code 1EJG.pdb) was selected as a test ground since it is a relatively small protein and its dynamics is widely studied. [11–13] The 46 residue Crambin consists of 657 atoms. Taking only the alpha carbons into account a set of 138 modes were obtained, six of which have a zero eigenvalue corresponding to the translation of the center of mass and three global rotational degrees of freedom around it.

All molecular dynamics simulations were performed for an N,P,T ensemble in explicit solvent (water) at 310 K using NAMD 2.5 package with CHARMM27 force field. The protein was solvated in a waterbox of 15 Å cushion and periodic boundary conditions were applied. Ions were added in order to represent a more typical biological environment. Langevin dynamics was used to control the system's temperature and pressure. All atoms were coupled to the heat bath of temperature of 310 K. A time step of 1fs was used. Nonbonded and electrostatic forces were evaluated at each time step. In order to keep all degrees of freedom no rigid bonds were used. All structures were translated so that their centers of mass are posit

we

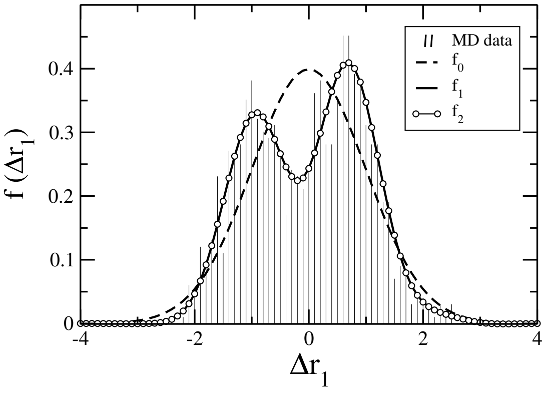

FIG. 3: A comparison of f_0 and f_1 and f_2 on the normalized histogram of the slowest mode. Note that f_1 and f_2 give the same marginal mode probabilities.

The full dataset consists of 8967 snapshots of 132 modes taken at 0.1 ps intervals. To prevent overfitting, every 9-th snapshot (a total of 996) was reserved as the test set and the rest (7971 snapshots) were used as the training set. Fig. 2 gives the time plots of a few of the sample modes.

In order to compare the quality of the three models f_0, f_1 and f_2 (given respectively by Eqs. (9, 10, and 12)), we use the average log likelihood of the snapshots in the test data, given the parameters optimized for the training data. Eq. (13) defines the average log likelihood of the data. $\mathbf{\Delta r}^{(i)}$ denotes the i-th snapshot and $\mathcal{N}$ is the number of snapshots.

$$\begin{aligned} \langle \log f(\mathbf{\Delta r}) \rangle &= \int f(\mathbf{\Delta r}) \log f(\mathbf{\Delta r}) d\mathbf{\Delta r} \\ &\approx \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \log f(\mathbf{\Delta r}^{(i)}) \qquad (13) \end{aligned}$$

From the definition of the log likelihood given in Eq. (13) it immediately follows that the entropy, S, can be estimated as $S/k_B = -\langle \log f(\mathbf{\Delta r}) \rangle$ where k_B is the Boltzmann constant. The average log likelihood based on f_0 is -187.5 per snapshot, corresponding to 115.5 kcal/mol contribution to the free energy. The latter is obtained as $TS = -RT \langle \log f(\mathbf{\Delta r}) \rangle$, $R = 1.986$ cal/mol, $T = 310$ K.

Fig. (3) compares the f_0 and f_1 distributions for the slowest mode, obtained from the test data. The anharmonicity of the dynamics is clear and is well represented by f_1. The free energy equivalent of the f_1 entropy is 115.1 kcal/mol which is only 0.4% less than that of f_0.

The f_2 approximation in Eq. (8) introduces pairwise interactions between modes. Note that, the correction due to the mode-coupling terms introduced in f_2 is invisible at the level of the marginal distibutions, such as in Fig. (3). Therefore we compare in the first two panels of Fig. 4 the contour plots of f_1 and f_2 distributions with the scatter plot of the two slowest modes. The free energy landscape of the protein is captured visibly better by f_2 when compared to f_1. The contribution at the level of f_2 to free energy equivalent of entropy is 110.9

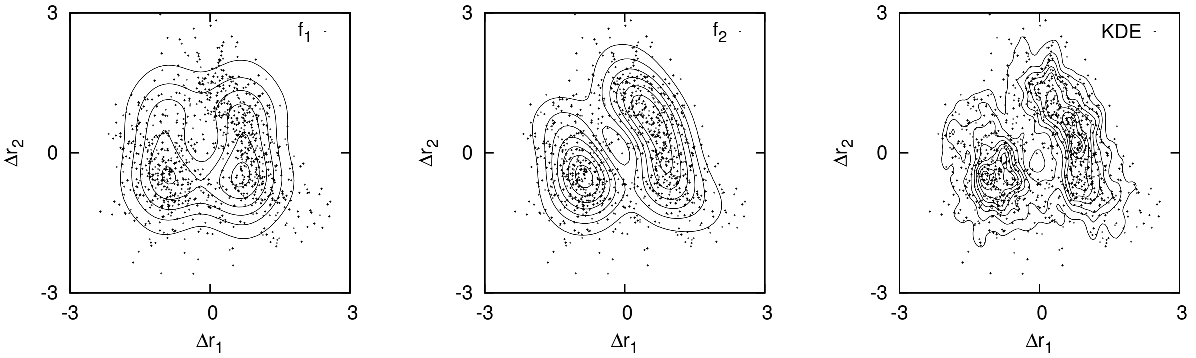

FIG. 4: A comparison of f_1, f_2 and KDE against a scatter plot of the two slowest modes.

kcal/mol. This results in a correction of ≈ -4.6 kcal/mol to the free energy *solely due to mode-coupling.*

Higher-order coupling

To quantify the effects of phenomena beyond pairwise interactions, we further used non-parametric kernel density estimation (KDE) procedure to model the probability density of the training set. In this approach, one fits the data using second-order Gaussian kernels with fixed bandwidths optimized by likelihood cross validation. We used the "npudensbw" and "npudens" procedures from the "np" package for the "R" statistical computing environment [14].

In Fig. 4, the results of KDE are presented in the third panel. These results are representative of all orders of contributions to anharmonicity and mode coupling and may be used to obtain a reasonable upper bound on the impact of mode-coupling on protein energetics. The free energy equivalent of the entropy calculated by the KDE is 108.4 kcal/mol. This shows that the total reduction in entropy due to anharmonicities and coupling relative to Gaussian is 7.1 kcal/mol, or 6.1% for Crambin.

In conclusion, the probability distributions of residue fluctutations obtained by the Hermite series expansion and the KDE give consistent measures of the fluctuational entropy of the protein Crambin in its native state. The Gaussian approximation f_0 gives a value of $TS = 115.5$ kcal/mol. Introduction of anharmonicities in the absence of mode coupling reduces the entropy to $TS = 115.1$ kcal/mol. Inclusion of second order mode coupling further reduces the entropy to $TS = 110.9$ kcal/mol. The KDE, which takes all orders of correlations and mode coupling into account yields an entropy of $TS = 108.4$ kcal/mol. In conclusion, we can state that although correlations introduce strong changes in the shape of the probability distribution their maximum effect on the entropy is only about 6% as determined from the difference between the Gaussian and the KDE approximations.

[1] I. Bahar, A. Atilgan, and B. Erman, Folding and Design **2**, 173 (1997).

[2] A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).

[3] O. Yogurtcu, M. Gur, and B. Erman, J. Chem. Phys. **130** (2009).

[4] S. Hayward, A. Kitao, and N. Go, Protein Science **3**, 936 (1994).

[5] F. Pontiggia, G. Colombo, C. Micheletti, and H. Orland, Phys. Rev. Lett. **98**, 48102 (2007).

[6] K. Moritsugu, O. Miyashita, and A. Kidera, Phys. Rev. Lett. **85**, 3970 (2000).

[7] D. Leitner (2008).

[8] A. García, Phys. Rev. Lett. **68**, 2696 (1992).

[9] P. Flory and D. Yoon, J. Chem. Phys. **61**, 5358 (1974).

[10] Q. Cui and I. Bahar, *Normal mode analysis: theory and applications to biological and chemical systems* (CRC Press, 2006).

[11] M. Teeter and D. Case, J. Phys. Chem. **94**, 8091 (1990).

[12] M. Levitt, C. Sander, and P. Stern, J. Mol. Biol. **181**, 423 (1985).

[13] O. Lange and H. Grubmuller, J. Phys. Chem. B **110**, 22842 (2006).

[14] T. Hayfield and J. Racine, Journal of Statistical Software **27**, 1 (2008).