Anharmonicity and mode-coupling in a fluctuating protein

We here develop a general framework for the residue fluctuations that simultaneously incorporates both anharmonicity and mode-coupling in a unified formalism. We show that both deviations from the pure Gaussian model are important for modeling the multidimensional energy landscape in the vicinity of the native state, even at physiological conditions where fluctuations are relatively small.

PACS numbers:

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]. However, it was recently shown that the slowest oscillatory modes of a protein are strongly anharmonic [2], 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 [3] and not captured by the harmonicity assumption.

Sampling the time evolution of a protein by using molecular dynamics reveals a multivariate probability distribution function $f(\mathbf{\Delta R})$ 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 p.d.f., where only C_α atoms are considered; i.e., N is the number of residues. Since the deviations from the free energy minimum should be harmonic for sufficiently small amplitudes, Hermite polynomials, which have a Gaussian kernel, constitute a natural basis for representing $f(\mathbf{\Delta R})$. First, following Ref. [2], we perform the transformation

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

which diagonalizes the covariance matrix $\Gamma \equiv \langle \mathbf{\Delta R \Delta R}^T \rangle$ and would give the normal modes of the protein if fluctuations were harmonic. Otherwise, the distribution function for $\{\Delta r\}$ in its most general form, can be expressed as [4]

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

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 in this "normal" basis 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 $\mathbf{C}_\nu = 0, \; \forall \nu$.

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

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

Above, $g(\mathbf{x}) = (2\pi)^{3N/2} \exp(-\mathbf{x}^2/2)$ is the multi-dimensional 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 $f(\mathbf{\Delta r})$ follow from the orthogonality relation as

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

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

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

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 $\mathbf{H}_\nu^{i_1 i_2 \cdots i_\nu}(\Delta r_k, \Delta r_l)$ 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 for the optimal cut-off rank $\nu = 16$ (see below), the number of distinct expectation values to be extracted from 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 $\mathbf{H}_\nu^{i_1 i_2 \cdots i_\nu}(\mathbf{\Delta r})$ 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,

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

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 $\{\mathbf{\Delta r}\}$ basis is diagonal further implies that

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

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

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

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

The first term in $[\cdots]$ corresponds to a purely harmonic given by the Gaussian probability distribution

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

This is the starting point for most of the past studies on protein fluctuations[**?**]. The next in Eq. (12) term is appreciable when the modes 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

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

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. (12) 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, these corrections are transparent to the marginal distributions

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

as a merit of the orthogonality relation in Eq. (**??**). Therefore, even if the marginal distributions are reproduced to good accuracy, the multi-dimensional 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 2CI2. To this end, we improve the approximation in Eq. (10) one step further and set $f(\mathbf{\Delta r})$ to

$$f_2(\mathbf{\Delta r}) = \frac{1}{\sqrt{(2\pi)^N}} e^{-\sum_i \Delta r_i^2/2} \Bigg[1 + \sum_i \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \Big\langle H_\nu(\Delta r_i) \Big\rangle H_\nu(\Delta r_i) + \sum_{i \neq j} \sum_{\nu=3}^{\infty} \frac{1}{\nu!} \sum_{p=1}^{\nu-1} \binom{\nu}{p} \Big\langle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \Big\rangle H_p(\Delta r_i) H_{\nu-p}(\Delta r_j) \Bigg] \tag{12}$$

which takes into account the mode-coupling corrections at the lowest order they appear and ignores higher-order terms cubic in H_ν.

Crambin (Protein Data bank code 1EJG.pdb) was selected as an test ground since it is a relatively small protein and its dynamics is widely studied [**? ? ?**].

The 46 residue Crambin consists of 657 atoms. Taking only the alpha carbons into account a set of 138 modes were obtained. Then the overall rotational and translational motions were eliminated since they are irrelevant for the internal motion [**?**]. All structures were translated so that their center of mass is positioned at the origin and rotated to obtain the best mass weighted RMSD fit with the initial structure.

The fluctuation $\mathbf{\Delta R}$ of atoms are defined by $\mathbf{\Delta R} = \mathbf{R} - \bar{\mathbf{R}}$, where $\bar{\mathbf{R}}$ are the mean atomic coordinates and

hence is a time independent quantity, which defines an average configuration obtained by the protein during the part of the trajectory that we use for calculations. The potential energy of this mean configuration is V_0. In the remaining part of the paper we characterize the deviations of the potential energy of the residues from that of the reference conformation.

The instantaneous fluctuations are transformed into modal space using Eq. (1) where the average denoted by the angular brackets is taken over the trajectory. [**?** **?**]. We let e represent the eigenvector matrix that diagonalizes $\Gamma^{-1/2} = \left\langle \mathbf{\Delta R \Delta R^T} \right\rangle^{-1/2}$, and λ represent the eigenvalues. Then, $\Gamma^{-1/2} = \text{diag}\ \lambda^{-1/2}\, e^T$ and the fluctuations $\mathbf{\Delta r}$ are the fluctuations in mode space spanned by the eigenvectors, e [**?**]. The components of the real trajectory that correspond to a given mode is obtained simply by keeping the eigenvalue of interest, equating all the others to zero, followed by a back transformation of Eq. (1).

The full dataset consists of 8967 snapshots of 132 modes. 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. Because of normalization, each mode has zero mean and unit variance. Figure 1 gives the time plots of a few of the sample modes.

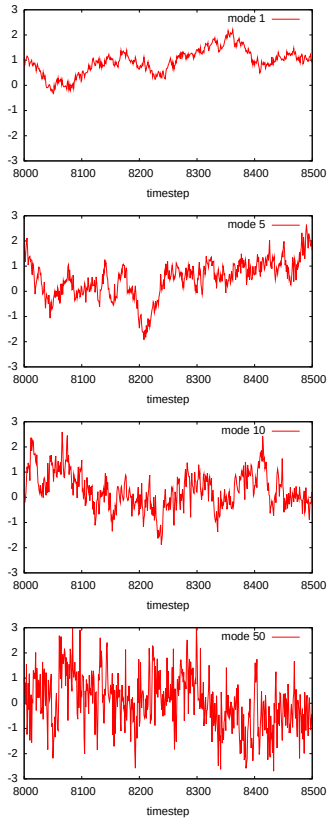

FIG. 1: Time plots of sample modes.

In order to compare the quality of the three models (f_0, f_1 and f_2 given respectively by Eq. (9, Eq. (10) and Eq. (12)), we use the average log likelihood of the snapshots in the test data given parameters optimized for the training data[1]. Eq. (13) defines the average log likelihood of the data. $\mathbf{\Delta r}^{(i)}$ denotes the i'th snapshot and N is the number of snapshots.

$$\langle \log f(\mathbf{\Delta r}) \rangle = \frac{1}{N} \sum_{i=1}^{N} \log f(\mathbf{\Delta r}^{(i)}) \qquad (13)$$

The average log likelihood based on f_0 is -187.449 per snapshot, corresponding to 111.5 kcal/mol contribution to the free energy at room temperature ($= \langle \log f \rangle k_B \times 1.6 \cdot 10^{-19} \times 6.02 \cdot 10^{23}/4200$ - Alkan). The same measurement on f_1 and f_2 needs more care, because the approximation may result in negative probabilities. Using the slowest 90(40) modes for $f_1(f_2)$ fixes the problem, except at a few data points which correspond to low probability regions of the configuration space and therefore can be safely ignored.

Figure 2 compares the f_0 and f_1 distributions with the test data histogram for mode 1. Although f_1 fits the marginal mode distributions better, modes are still assumed independent of each other.

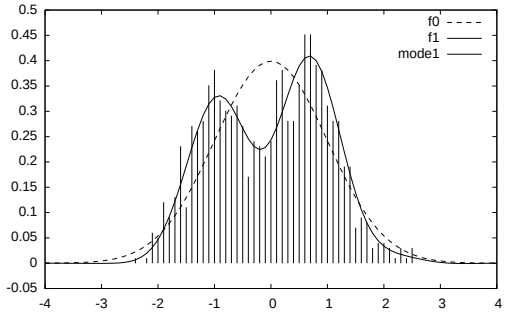

FIG. 2: A comparison of f_0 and f_1 and f_2 with the normalized histogram for mode 1. Note that f_1 and f_2 overlap at this level of marginal mode probabilities.

The f_2 approximation in Eq. (12) can model pairwise interactions between modes. Considering the slowest 40 modes, f_2 results in a correction of ≈ 0.9 kcal/mol to the free energy *solely due to mode-coupling.*

f_2 achieves a significantly better fit to the data compared to f_1. The two models give identical marginal distributions to individual modes. The difference comes from f_2's better modeling of pairwise interactions. Figure 3 compares f_1 and f_2 distributions with the scatter plot of the first two modes.

[1] except for f_0, which has no parameters to optimize

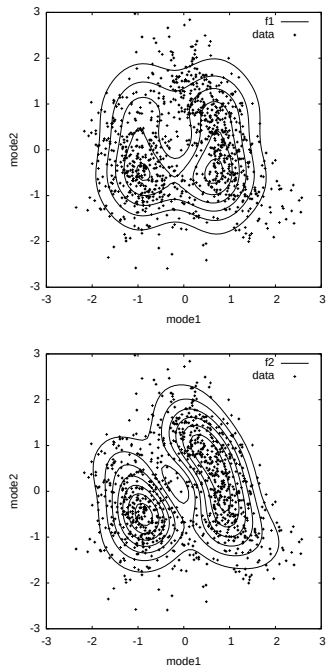

FIG. 3: A comparison of f_1 and f_2 against a scatter plot of the first two modes.

[1] Reference GNM, ENM, etc.
[2] Yogurtcu *et al.* (2009).
[3] Reference on mode coupling, enegry transfer between modes, etc.
[4] Flory, 1976.