Formal Description of Covariance Analysis and Collection of Intermediate Structural Ensembles

The native-centric topological modeling program outputs a trajectory of particle positions and energies. Let us represent the residue energy data as a vector function $\vec{E}(t)$ and define a basis $\{\widehat{e}_i\}$ $(i = \{1, \ldots, N_{\text{res}}\})$ where $\widehat{e}_i$ is a unit vector for the energy of residue i and N_{res} is the number of residues in the simulated protein. In this basis, each component of $\vec{E}(t)$ gives the energy of a single residue at simulation sampling period t. Call these components $e_i(t)$. Thus,

$$\vec{E}(t) = \sum_{i=1}^{N_{\text{res}}} e_i(t)\widehat{e}_i \qquad t = \{1, \ldots, N_{\text{samp}}\},$$

where N_{samp} is the number of sampling periods. Let us similarly represent the residue position data as a vector function $\vec{Q}(t)$. Define a basis $\{\widehat{q}_j\}$ $(j = \{1, \ldots, 3N_{\text{res}}\})$ consisting of three Cartesian unit position vectors for each particle, and let $q_j(t)$ be the component of $\vec{Q}(t)$ along $\widehat{q}_j$. Thus,

$$\vec{Q}(t) = \sum_{j=1}^{3N_{\text{res}}} q_j(t)\widehat{q}_j$$

Now let us define the covariance matrix $\widetilde{C}$. Element C_{ij} of $\widetilde{C}$ is the covariance of $e_i(t)$ and $e_j(t)$:

$$C_{ij} = \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} \left(e_i(t) - \bar{e}_i\right)\left(e_j(t) - \bar{e}_j\right), \qquad \text{where} \qquad \bar{e}_i = \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} e_i(t)$$

It is well known that every symmetric matrix is orthogonally diagonalizable. Therefore $\widetilde{C}$ is always diagonalizable and its eigenvectors are always mutually orthogonal. Let $\widetilde{U}$ be the unitary transformation that diagonalizes $\widetilde{C}$, let λ_i $(i = \{1, \ldots, N_{\text{res}}\})$ be the eigenvalues of $\widetilde{C}$, and let $\widehat{v}_i$ be corresponding normalized eigenvectors. Then

$$\left(\widetilde{U}\widetilde{C}\widetilde{U}^{\dagger}\right)_{ij} = \lambda_i \delta_{ij}, \qquad \text{where} \qquad \widetilde{U} = \sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} \left(\widehat{v}_j \cdot \widehat{e}_i\right) \widehat{v}_j \widehat{e}_i^{\dagger}$$

The eigenvectors $\widehat{v}_i$ define the system's principal energy modes. The original residue energy data can be projected onto these modes by applying the transformation $\widetilde{U}$ to $\vec{E}(t)$:

$$\widetilde{U}\vec{E}(t) = \left(\sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} \left(\widehat{v}_j \cdot \widehat{e}_i\right) \widehat{v}_j \widehat{e}_i^{\dagger} \right) \left(\sum_{k=1}^{N_{\text{res}}} e_k(t) \widehat{e}_k \right) = \sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} e_i(t) \left(\widehat{v}_j \cdot \widehat{e}_i\right) \widehat{v}_j$$

Therefore,

$$\widetilde{U}\vec{E}(t) = \sum_{j=1}^{N_{\text{res}}} \varepsilon_j(t)\widehat{v}_j \qquad \text{where} \qquad \varepsilon_j(t) = \sum_{i=1}^{N_{\text{res}}} e_i(t)\left(\widehat{v}_j \cdot \widehat{e}_i\right)$$

Component $\varepsilon_i(t)$ of $\widetilde{U}\vec{E}(t)$ is the projection of the energy at time t along mode $\widehat{v}_i$.

We now wish to collect structural data from simulation time points t for which the energy projection $\varepsilon_i(t)$ along some selected mode $\widehat{v}_i$ is extremal. Define structural ensembles

$$S_{i+} = \{\vec{Q}(t) | \varepsilon_i(t) \geq \tau_{i+}\} \qquad \text{and} \qquad S_{i-} = \{\vec{Q}(t) | \varepsilon_i(t) \leq \tau_{i-}\}$$

where S_{i+} corresponds to maximal values of $\varepsilon_i(t)$, S_{i-} corresponds to minimal values of $\varepsilon_i(t)$, and the threshold values $\tau_{i\pm}$ are chosen such that $|S_{i-}| = |S_{i+}| = 0.01 \cdot N_{\text{samp}}$.

The structures in each ensemble $S_{i\pm}$ are randomly oriented and must next be aligned with respect to an external reference structure. Define the reference structure as a vector $\vec{N}$ with components $\{n_j\}$ $(j = 1, \ldots, 3N_{\text{res}})$ in the same basis as $\vec{Q}(t)$:

$$\vec{N} = \sum_{j=1}^{3N_{\text{res}}} n_j \widehat{q}_j$$

Some subset of the system's particle coordinates $j = j_{\text{start}}, \ldots, j_{\text{end}}$ is selected, typically corresponding to particles in a particular secondary structure element of the simulated protein. Each structure $\vec{Q}(t) \in S_{i\pm}$ is rotated to obtain a new structure $\vec{Q}'(t)$ with a minimal mean square difference from $\vec{N}$ in the coordinates $j = j_{\text{start}}, \ldots, j_{\text{end}}$ using the algorithm of Kabsch (1). The rotated structures form a new ensemble $S'_{i\pm}$:

$$S'_{i\pm} = \{\vec{Q}'(t) | \vec{Q}(t) \in S_{i\pm}\} \qquad \text{where} \qquad \vec{Q}'(t) = R_{\vec{Q}(t),\vec{N}}\vec{Q}(t) = \sum_{j=1}^{3N_{\text{res}}} q'_j \widehat{q}_j$$

Here, $R_{\vec{Q}(t),\vec{N}}$ is the proper rotation matrix that minimizes $\alpha_{\vec{Q}(t),\vec{N}}$:

$$\alpha_{\vec{Q}(t),\vec{N}} = \sum_{j=j_{\text{start}}}^{j_{\text{end}}} \left(q'_i - n_i\right)^2$$

The ensemble $S'_{i\pm}$ has thousands of members and is difficult to visualize. Let us construct a reduced ensemble $s_{i\pm}$, a representative subset of $S'_{i\pm}$ with only 50 members:

$$s_{i\pm} \subset S'_{i\pm} \qquad |s_{i\pm}| = 50$$

The elements of $s_{i\pm}$ are chosen by a Monte Carlo algorithm to minimize β:

$$\beta = \|\vec{\mu}(s_{i\pm}) - \vec{\mu}(S'_{i\pm})\|^2 + \|\vec{\sigma}(s_{i\pm}) - \vec{\sigma}(S'_{i\pm})\|^2,$$

where $\vec{\mu}(A)$ is a vector of average coordinates for some ensemble of structures A:

$$A = \{\vec{Q}_A\} \qquad\qquad \vec{\mu}(A) = \frac{1}{|A|}\sum_A \vec{Q}_A$$

and $\vec{\sigma}(A)$ is a vector of coordinate standard deviations:

$$\vec{\sigma}(A) = \sum_{j=1}^{3N_{\text{res}}} \left(\frac{1}{|A|} \sum_A \left(q_{j,A} - \bar{q}_{j,A}\right)^2 \right)^{\frac{1}{2}} \widehat{q}_j$$

where

$$\vec{Q}_A = \sum_{j=1}^{3N_{\text{res}}} q_{j,A}\widehat{q}_j \qquad\qquad \text{and} \qquad\qquad \bar{q}_{j,A} = \frac{1}{A}\sum_A q_{j,A}$$

Theoretical Justification of Covariance Analysis

This discussion is based on the work of García (2).

Consider a simulation energy trajectory. The trajectory data can be described as a vector $\vec{E}(t)$, where each component of the vector describes the energy of a different particle during sampling period t ($t = \{1, \ldots, N_{\text{samp}}\}$). Each component fluctuates, and the fluctuations of different components are not independent. Our objective is to find an efficient description of these correlated fluctuations.

At each time point we can calculate the displacement of $\vec{E}(t)$ from its mean position. Let us find the unit vector $\widehat{v}$ that gives the most probable direction of this displacement. To do this, we will find $\widehat{v}$ such that the mean square projection of the displacement along $\widehat{v}$ is maximal. This is equivalent to maximizing the following function $f(\widehat{v})$ under the normalization constraint $\widehat{v} \cdot \widehat{v} = 1$:

$$f(\widehat{v}) = \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} \left[\left(\vec{E}(t) - \left\langle \vec{E} \right\rangle \right) \cdot \widehat{v} \right]^2 \qquad \text{where} \qquad \left\langle \vec{E} \right\rangle = \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} \vec{E}(t) \qquad [\mathbf{1}]$$

It is useful to rearrange our expression for $f(\widehat{v})$ in the following way:

$$\begin{aligned} f(\widehat{v}) &= \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} \left[\widehat{v} \cdot \left(\vec{E}(t) - \left\langle \vec{E} \right\rangle\right)\right] \left[\left(\vec{E}(t) - \left\langle \vec{E} \right\rangle\right) \cdot \widehat{v}\right] \\ &= \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} \widehat{v}^{\dagger} \left(\vec{E}(t) - \left\langle \vec{E} \right\rangle\right) \left(\vec{E}(t) - \left\langle \vec{E} \right\rangle\right)^{\dagger} \widehat{v} \\ &= \widehat{v}^{\dagger} \widetilde{C} \widehat{v} \\ &= \left(\widetilde{C}\widehat{v}\right) \cdot \widehat{v}, \end{aligned}$$

where we have introduced the covariance matrix $\widetilde{C}$:

$$\widetilde{C} = \frac{1}{N_{\text{samp}}} \sum_{t=1}^{N_{\text{samp}}} \left(\vec{E}(t) - \left\langle \vec{E} \right\rangle\right) \left(\vec{E}(t) - \left\langle \vec{E} \right\rangle\right)^{\dagger}$$

Note that $\widetilde{C}$ is a symmetric matrix.

Following the method of Lagrange multipliers, the critical points of $f(\widehat{v})$ under the normalization constraint correspond exactly to the critical points of the following Lagrangian function $g(\widehat{v}, \lambda)$:

$$\begin{aligned} g(\widehat{v}, \lambda) &= f(\widehat{v}) - \lambda\left(\widehat{v} \cdot \widehat{v} - 1\right) \\ &= \left(\widetilde{C}\widehat{v}\right) \cdot \widehat{v} - \lambda\left(\widehat{v} \cdot \widehat{v} - 1\right) \end{aligned}$$

We will find the critical points of $g(\widehat{v}, \lambda)$ by finding $\widehat{v}$ such that $\nabla g = 0$.

At this point, it is convenient to expand $\widetilde{C}$ and $\widehat{v}$ in terms of components by using the following substitutions:

$$\widetilde{C} = \sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} C_{ij} \widehat{e}_i \widehat{e}_j^{\dagger} \qquad \widehat{v} = \sum_{i=1}^{N_{\text{res}}} v_i \widehat{e}_i \tag{\textbf{2}}$$

Note that since $\widetilde{C}$ is symmetric, $C_{ij} = C_{ji}$. It is also useful to evaluate the product $\widetilde{C}\widehat{v}$, which appears in $g(\widehat{v}, \lambda)$ and will also be useful later.

$$\widetilde{C}\widehat{v} = \left(\sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} C_{ij} \widehat{e}_i \widehat{e}_j^{\dagger}\right) \left(\sum_{n=1}^{N_{\text{res}}} v_n \widehat{e}_n\right) = \sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} C_{ij} v_j \widehat{e}_i \tag{\textbf{3}}$$

Making these substitutions in our expression for $g(\widehat{v}, \lambda)$ gives

$$\begin{aligned} g(\widehat{v}, \lambda) &= \left(\widetilde{C}\widehat{v}\right) \cdot \widehat{v} - \lambda \left(\widehat{v} \cdot \widehat{v} - 1\right) \\ &= \left(\sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} C_{ij} v_j \widehat{e}_i\right) \cdot \left(\sum_{n=1}^{N_{\text{res}}} v_n \widehat{e}_n\right) - \lambda \left(\sum_{i=1}^{N_{\text{res}}} v_i^2 - 1\right) \\ &= \sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} C_{ij} v_i v_j - \lambda \left(\sum_{i=1}^{N_{\text{res}}} v_i^2 - 1\right) \end{aligned}$$

Now let us evaluate ∇g. For clarity we will not consider λ as a variable when evaluating the gradient. Note that evaluating $\partial g(\widehat{v}, \lambda)/\partial \lambda = 0$ simply returns the normalization condition, so we don't lose any information by evaluating ∇g in this way.

$$\begin{aligned} \nabla g &= \sum_{k=1}^{N_{\text{res}}} \widehat{e}_k \frac{\partial}{\partial v_k} g(\widehat{v}, \lambda) \\ &= \sum_{k=1}^{N_{\text{res}}} \widehat{e}_k \left[\frac{\partial}{\partial v_k} \sum_{i=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} C_{ij} v_i v_j - \lambda \frac{\partial}{\partial v_k} \sum_{i=1}^{N_{\text{res}}} v_i^2\right] \end{aligned}$$

To evaluate the derivative of the sums, let us drop all terms that do not involve v_k:

$$\begin{aligned} \nabla g &= \sum_{k=1}^{N_{\text{res}}} \widehat{e}_k \left[\frac{\partial}{\partial v_k} \left(\sum_{i=1, i\neq k}^{N_{\text{res}}} C_{ik} v_i v_k + \sum_{j=1, j\neq k}^{N_{\text{res}}} C_{kj} v_k v_j + C_{kk} v_k^2\right) - 2\lambda v_k\right] \\ &= \sum_{k=1}^{N_{\text{res}}} \widehat{e}_k \left(\sum_{i=1, i\neq k}^{N_{\text{res}}} C_{ik} v_i + \sum_{j=1, j\neq k}^{N_{\text{res}}} C_{kj} v_j + 2C_{kk} v_k - 2\lambda v_k\right) \end{aligned}$$

Using the symmetry relation $C_{ij} = C_{ji}$ to combine summations:

$$\nabla g = 2 \sum_{k=1}^{N_{\text{res}}} \sum_{j=1}^{N_{\text{res}}} C_{kj} v_j \widehat{e}_k - 2\lambda \sum_{k=1}^{N_{\text{res}}} v_k \widehat{e}_k$$

Now, using Eqs. **2** and **3**, we get:

$$\nabla g = 2\widetilde{C}\widehat{v} - 2\lambda \widehat{v}$$

We see that, to satisfy $\nabla g = 0$, $\widehat{v}$ must be an eigenvector of $\widetilde{C}$.

$$\widetilde{C}\widehat{v} = \lambda \widehat{v}$$

It is well known that symmetric matrices such as $\widetilde{C}$ are orthogonally diagonalizable. Therefore, some orthonormal set of eigenvectors $\{\widehat{v}_k\}$ exists with corresponding eigenvalues$\{\lambda_k\}$ $(k = \{1, \ldots, N_{\text{res}}\})$. Let us index the eigenvectors and eigenvalues in order of decreasing magnitude, so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{N_{\text{res}}}$.

The eigenvectors $\{\widehat{v}_k\}$ are the local maxima of $f(\widehat{v})$ subject to the normalization condition. The global maximum can be found by substituting each $\widehat{v}_k$ into $f(\widehat{v})$.

$$f(\widehat{v}_k) = \left(\widetilde{C}\widehat{v}_k\right)\cdot\widehat{v}_k = \lambda_k\widehat{v}_k\cdot\widehat{v}_k = \lambda_k \qquad \mathbf{[4]}$$

Therefore, the global maximum is $f(\widehat{v}_1)$, because λ_1 is the greatest of the eigenvalues.

We see that $\widehat{v}_1$ is the vector which best describes the fluctuations in $\vec{E}(t)$. The remaining eigenvectors describe directions orthogonal in energy space to $\widehat{v}_1$ that also maximize $f(\widehat{v})$. For example, if the original data $\vec{E}(t)$ were treated to remove fluctuations along $\widehat{v}_1$, eigenvector $\widehat{v}_2$ would be the best description of the remaining fluctuations.

Eqs. **1** and **4** establish that the eigenvalue λ_i gives the mean square displacement of $\vec{E}(t)$ from its mean position along the direction given by $\widehat{v}_i$. The total mean square displacement is therefore given by $\sum_{i=1}^{N_{\mathrm{res}}} \lambda_i$, and $\lambda_j / \sum_{i=1}^{N_{\mathrm{res}}} \lambda_i$ gives the fraction of the mean square displacement accounted for by fluctuations along $\widehat{v}_j$.

Convergence Test for Covariance Analysis Results

Covariance analysis results were tested for convergence by using the following method, which is based on that of Amadei *et al.* (3). A series of data sets is prepared by truncation of the full set, each containing data from $n \times 25,000$ simulation periods where $n = 1, 2, \ldots, 9$ and a covariance analysis is performed on each set. The resulting covariance matrix eigenvectors are compared by using the formula $S(n) = \frac{1}{10}\sum_{i=1}^{10}\left(\widehat{v}_{i,n} \cdot \widehat{v}_{i,n+1}\right)^2$, where $\widehat{v}_{i,n}$ is eigenvector i from the dataset with $n \times 25,000$ samples. For each data set, only the 10 eigenvectors with the largest eigenvalues are considered; the remaining eigenvectors are never interpreted so their convergence is unimportant. We consider the covariance analysis results to be well converged if $S(n)$ is stable and close to unity for the four or five largest values of n, which indicates that the addition of more data would have little effect. All covariance analysis results presented in this work are well converged.

1. Kabsch, W. (1978) *Acta Crystallogr. A* **34,** 827–828.
2. García, A. E. (1992) *Phys. Rev. Lett.* **68,** 2696–2699.
3. Amadei, A., Ceruso, M. A. & Di Nola, A. (1999) *Proteins Struct. Funct. Genet.* **36,** 419–424.