1-27-2018

# Comparisons of Protein Dynamics from Experimental Structure Ensembles, Molecular Dynamics Ensembles, and Coarse-Grained Elastic Network Models

Kannan Sankar
*Iowa State University*

Sambit K. Mishra
*Iowa State University*

Robert L. Jernigan
*Iowa State University*, jernigan@iastate.edu

# Comparisons of Protein Dynamics from Experimental Structure Ensembles, Molecular Dynamics Ensembles, and Coarse-Grained Elastic Network Models

**Abstract**

Predicting protein motions is important for bridging the gap between protein structure and function. With growing numbers of structures of the same, or closely related proteins becoming available, it is now possible to understand more about the intrinsic dynamics of a protein with principal component analysis (PCA) of the motions apparent within ensembles of experimental structures. In this paper, we compare the motions extracted from experimental ensembles of 50 different proteins with the modes of motion predicted by several types of coarse-grained elastic network models (ENMs) which additionally take into account more details of either the protein geometry or the amino acid specificity. We further compare the structural variations in the experimental ensembles with the motions sampled in molecular dynamics (MD) simulations for a smaller subset of 17 proteins with available trajectories. We find that the correlations between the motions extracted from MD trajectories and experimental structure ensembles are slightly different than for the ENMs, possibly reflecting potential sampling biases. We find that there are small gains in the predictive power of the ENMs in reproducing motions present in either experimental or MD ensembles by accounting for the protein geometry rather than the amino acid specificity of the interactions.

**Disciplines**

Biochemistry | Bioinformatics | Biophysics | Molecular Biology | Structural Biology

# Comparisons among Protein Dynamics from Experimental Structure Ensembles, Molecular Dynamics Ensembles, and Coarse-Grained Elastic Network Models

*Kannan Sankar [†‡§], Sambit K. Mishra[†‡] and Robert L. Jernigan[†‡*]*

[†]Bioinformatics and Computational Biology Interdepartmental Graduate Program, Iowa State University, Ames, IA 50011-1178, USA

[‡]Roy J. Carver Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA 50011-1178, USA

# ABSTRACT

Predicting protein motions is important for bridging the gap between protein structure and function. With growing numbers of structures of the same, or closely related proteins becoming available, it is now possible to understand more about the intrinsic dynamics of a protein with principal component analysis (PCA) of the motions apparent within ensembles of experimental structures. In this paper, we compare the motions extracted from experimental ensembles of 50 different proteins with the modes of motion predicted by several types of coarse-grained elastic network models (ENMs) which additionally take into account more details of either the protein geometry or the amino acid specificity. We further compare the structural variations in the experimental ensembles with the motions sampled in molecular dynamics (MD) simulations for a smaller subset of 17 proteins with available trajectories. We find that the correlations between the motions extracted from MD trajectories and experimental structure ensembles are slightly different than for the ENMs, possibly reflecting potential sampling biases. We find that there are small gains in the predictive power of the ENMs in reproducing motions present in either experimental or MD ensembles by accounting for the protein geometry rather than the amino acid specificity of the interactions.

## INTRODUCTION

Predicting conformational changes in proteins has long been a topic of interest to many who aim to understand protein function and mechanism. Multiple structures of the same protein, or closely related proteins,  have been solved by different experimental methods - X-ray crystallography[1], NMR spectroscopy[2] and more recently by cryo-electron microscopy,[3] under different conditions, in the presence of different ligands, or of mutated protein. These techniques reveal information about the intrinsic protein dynamics. The set of essential motions accessible to a protein can be readily obtained by applying principal component analysis (PCA)[4] to the position coordinates of the aligned set of multiple experimental structures.[5–10]

Information on protein motions can also be obtained from computational simulations such as molecular dynamics (MD) or Monte Carlo (MC). However, these applications require significant computer resources, and do not always fully sample the entire conformational space accessible to a protein. Coarse-grained elastic network models (ENMs)[11–13] on the other hand, offer a faster and cheaper alternative to MD or MC simulations for sampling the intrinsic motions accessible to a protein. By modeling the protein as a string of beads (usually the $C^{\alpha}$ atoms) connected by harmonic springs (interactions), they are often able to capture the most important global motions. ENMs have been used extensively to study the intrinsic dynamics of a variety of biomolecules ranging from small globular and membrane proteins[14] to nucleic acids,[15] and even large biomolecular assemblies such as the ribosome[16–18] and GroEL.[19,20] They have been shown to accurately predict the crystallographic B-factors of diverse proteins[21,22] as well as to capture conformational changes between pairs of structures of the same protein.[23,24] The normal modes from ENMs have also been shown to capture structural variations extracted from multiple experimental structures of the same protein[25–27] or RNA.[28]

Specifically, here we focus on ENMs that provide the changes in the geometry, the Anisotropic Network Models (ANM).[29] A subject of some importance has been how to improve ENMs by accounting for either more specific details of protein geometry or the chemical nature of amino acids.[30,31] Hamacher and McCammon have shown that an extended ANM (**eANM**)[32] with spring constants based on the values of the Miyazawa-Jernigan (MJ) potential amino acid interaction energies[33] to account for the amino acid specificity of fluctuations performs better in reproducing crystallographic B-factors. We have also shown that the ANM can be significantly improved by weighting the spring constants between residues by the inverse powers of the distance of separation between them,[34] a model referred to as the parameter-free ANM (**pfANM**) (pf means that there is no cutoff parameter as in the traditional ANM). Other ways of adjusting the springs in ENMs are to use information from the variance-covariance matrix of position coordinates[35] or the mean square distance fluctuations[36] between residues from MD trajectory ensembles of the protein. We and others have also shown that using spring constants based on the variance of internal distance changes between residues also provides significant gains in the ability to reproduce experimentally observed conformational changes.[37,38]

In this work, we also introduce a modified version of Hamacher and McCammon's extended ANM (called **ccANM**) in which the spring constants between residues are based on the relative entropies of amino acid pairs rather than the relative energies of the pairs. This is based on our recent work, where we extracted a scale of relative entropies between amino acid pairs[39] based on the frequencies of contact changes between amino acid types during conformational changes within a dataset of proteins. This entropy measure yields significant gains in identifying native structures among decoy sets. Since these entropies measure the tendency for amino acid contacts to change, we hypothesize that information on relative entropies of the amino acid pairs might be more useful than their relative energies for differentiating among springs representing the interactions.

First, we systematically test the effectiveness of the classical coarse-grained ANM and four different variants of the ANM (that incorporate additional information either regarding protein geometry or amino acid specificity) in capturing the motions present in experimental structure ensembles of 50 different proteins. In addition, for a smaller subset of 17 proteins where MD trajectories are available, we also compare the motions present in the experimental ensembles to those in the MD ensembles. Our results suggest that the protein motions as extracted from experimental ensembles can differ significantly from those obtained through MD simulations. Whether this reflects the difference between the crystal environments and the simulation conditions, or a failure of simulations to fully capture the characteristic dynamics remains an open question. In addition, we also investigate how well the motions present in either the experimental or MD ensembles are captured by a variety of simple coarse-grained elastic network models.

## METHODS

**Experimental Structure Ensemble Data.** A set of experimental structure ensembles for 50 different proteins (Table S1) were collected in our previous work,[40] which we are utilizing here. We refer the reader to this previous work for the list of structures in each ensemble set. These structures were obtained by a clustering of the Protein Data Bank (PDB)[41] at the 95% sequence identity level. “Only the monomeric structures are retained. The structures in each cluster were aligned using the multiple structure alignment program MUSTANG,[42] and the corresponding structure-based sequence alignment was used as a guide to remove any residues and/or structures that introduced significant gaps in the middle of the alignment (relatively few such cases). The final set of aligned structures from our previous work has been used for the experimental protein ensembles. For construction of ANMs, the structure with the lowest average root mean square deviation (RMSD) from all other structures is chosen as the representative structure for each ensemble (see Table S1 for the list of these representative structures). The distributions of the average RMSDs in each ensemble can be found in our previous work.[40]

**Molecular Dynamics Trajectories.** For each experimental protein set, we have searched for homologous entries in the MoDEL database,[43] a repository of publicly available MD trajectories. Since the set of proteins in each cluster have a high sequence identity (≥ 95%), we choose a protein randomly from each cluster and search for its homologs. We set a threshold on the sequence identity of 35% for this selection. For clusters with multiple available homologs, we only choose the one with the highest sequence identity. We then download the $C^\alpha$ atom trajectories for the selected homologs for each cluster from the MoDEL database. A list of the proteins whose trajectories were used is given in Table S2.

In order to obtain a common reference frame, we transform the coordinates of the MD trajectories from their native frame to the frame of their experimental homologs. We do this by superimposing the first frame from each MD trajectory onto the representative structure from the corresponding experimental ensemble set; and then superimposing all the other frames onto the first frame. In order to identify a common subset of residues between the experimental and MD datasets, we then align the sequence of each MD homolog to the profile alignment of its respective experimental set (with *ClustalOmega)*[44] and retain only the subset of residues from the PDB structure in common with the MD homolog and the experimental ensemble. For generating ANMs, the starting PDB structure of each MD dataset is used as the representative of the ensemble (see Table S2).

**Principal Component Analysis of structural ensembles.** Information about protein dynamics is extracted from either the experimental ensemble or the MD trajectory ensemble by using PCA of the aligned set of structures (to remove rigid body motions). In each case, the dataset for PCA is a matrix $X_{n \times 3N}$ consisting of the X-, Y- and Z-coordinates of the $C^{\alpha}$ atoms of each of $N$ residues in the aligned set of $n$ structures. The variance-covariance matrix $\boldsymbol{C}_{3N \times 3N}$ of the position coordinates is constructed with its elements obtained as

$$C_{ij} = \langle X_{ij} - \langle X_i \rangle \rangle \langle X_{ij} - \langle X_j \rangle \rangle\,; \tag{1}$$

where the brackets refer to averages across all $n$ structures. Eigen-decomposition of the matrix $\boldsymbol{C}$ results in the eigenvectors, which are a set of orthogonal directions of the variations present in the dataset having corresponding eigenvalues denoting the variance along the corresponding directions. The principal component (PC) scores are obtained directly from the projections of the mean centered data points along these eigenvectors. The PCs are sorted in decreasing order of the corresponding eigenvalues and referred to as PC1, PC2, PC3 and so on, with PC1 capturing the most significant part of the structural variations.

**Coarse-grained elastic network models.** Next, we describe the various coarse-grained bead-spring models that we have used in our comparisons. Collectively these are all termed elastic network models (ENMs).

***Anisotropic Network Model (ANM).*** The ANM[29] is an elastic-network (bead-spring) model in which the $C^{\alpha}$ atoms of each residue in the protein are represented as beads and all interactions between residues are modeled as harmonic springs. Interactions between beads are usually restricted to physically close residues within a fixed distance cutoff $R_c$. There are two parameters in ANM: the distance cutoff $R_c$ and the spring constant $\gamma_{ij}$ between every pair of residues $i$ and $j$. Throughout this study, the value of $R_c$ has been set to 13 Å. In a classical ANM, all springs are assigned uniform values. In other words, for a protein with $N$ residues,

$$\gamma_{ij} = \gamma \ \forall \ i, j \ \epsilon \ \{1, \dots, N\} \tag{2}$$

All the springs are assumed to be in equilibrium in the starting structure and the potential energy $V$ of the system is computed as

$$V = \tfrac{1}{2} \textstyle\sum_{i,j=1}^{N} \gamma_{ij} \left(R_{ij} - R_{ij}^{0}\right)^2, \tag{3}$$

where $R_{ij}$ refers to the instantaneous displacement between atoms $i$ and $j$ and $R_{ij}^{0}$ refers to their equilibrium displacement. The Hessian matrix $\boldsymbol{H}$ of the system, with $N \times N$ superelements $H_{ij}$ is calculated as the matrix of second derivatives of the potential with respect to the Cartesian coordinate positions of the residues as

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 V}{\partial X_i \partial X_j} & \frac{\partial^2 V}{\partial X_i \partial Y_j} & \frac{\partial^2 V}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Y_i \partial X_j} & \frac{\partial^2 V}{\partial Y_i \partial Y_j} & \frac{\partial^2 V}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Z_i \partial X_j} & \frac{\partial^2 V}{\partial Z_i \partial Y_j} & \frac{\partial^2 V}{\partial Z_i \partial Z_j} \end{bmatrix} \tag{4}$$

The normal modes of motion from ANM are obtained as eigenvectors of the matrix $\boldsymbol{H}$; with the corresponding eigenvalues representing the square of frequencies of the modes. The correlations in motion between the residues along the X, Y and Z directions can be obtained from the corresponding super-elements of $\boldsymbol{H}^{-1}$ and the mean square fluctuations of each residue $i$ from the diagonal elements of the corresponding superelement $H_{ii}^{-1}$ as follows:

$$\langle \Delta R_i^2 \rangle = \frac{k_B T}{\gamma} trace(H_{ii}^{-1}) \tag{5}$$

The theoretical B-factors from the ANM can be conveniently calculated from the mean square fluctuations as

$$B_i^{calc} = 8\pi^2 \langle \Delta R_i^2 \rangle / 3 \tag{6}$$

In addition to the classical ANM, we also explore some different variants of the ANM. The basic idea of each of the modified ANMs is the same, with the only change being that the spring constants are modified somehow.

***Parameter-free ANM (pfANM).*** In the pfANM[34], one of the parameters, the $R_c$ is eliminated by allowing all residues to be connected, but instead of uniform springs the spring constants are taken to be proportional to a given inverse power $p$ of the distance $r_{ij}$ between them as in Eq. 7. Previously we found that $p = 6$ gave the best representation of the collective motions; whereas $p = 2$ best fit the experimental B-factors.[34]

$$\gamma_{ij} = \frac{1}{r_{ij}^{p}} \ \forall\ i,j \, \epsilon \, \{1,\dots,N\} \tag{7}$$

***Extended ANM (eANM).*** We use a simplified version of a modified ANM introduced by Hamacher and McCammon[32] in which the spring constants between a pair of non-adjacent contacting residues (as identified by $R_c$) is weighted by the absolute value of the Miyazawa-Jernigan (MJ) potential[33] energy $\left|\kappa_{ij}\right|$ between them. The spring stiffness between adjacent residues is set to a much larger value, $K$ = 82 RT/Å$^2$ in accordance with the values found for peptide bonds. That is,

$$\gamma_{ij} = \begin{cases} K \quad \text{if} \ \ |i-j| = 1 \\ 2\left|\kappa_{ij}\right| \ \ \text{if} \ \ |i-j| \neq 1 \text{ and } r_{ij} \leq R_c \end{cases} \forall\ i,j \, \epsilon \, \{1,\dots,N\} \tag{8}$$

***Contact-change based ANM (ccANM).*** This is a model similar to the eANM; except that the springs between non-adjacent contacting residues falling within the cutoff distance $R_c$ are weighted by the inverse of the contact-change based entropies[39] $s_{ij}$ between the amino acid pair. That is,

$$\gamma_{ij} = \begin{cases} K \quad if \ \ |i-j| = 1 \\ \frac{1}{s_{ij}} \quad if \ \ |i-j| \neq 1 \ \text{ and } r_{ij} \leq R_c \end{cases} \ \forall\ \ i,j \, \epsilon \, \{1,\dots,N\} \tag{9}$$

***Distance change based ANM (dcANM).*** This model captures internal distance-changes as observed within an ensemble of structures. For this variant of the ANM, the spring constants between each pair of residues is taken as the inverse of the variance of internal distances ($\sigma_{r_{ij}}^{2}$) between the residue pair over the set of structures (these spring constant values were further normalized such that they range between 0 and 1). In other words,

$$\gamma_{ij} = \frac{1}{\sigma^2_{r_{ij}}} \quad \forall \;\; i, j \, \epsilon \, \{1, \dots, N\} \tag{10}$$

**Performance Evaluation of the ENMs.** We measure the performance of each ENM in terms of how well it can reproduce the protein structural variations present within an ensemble. The directions of motions from the ENM are obtained directly from the ENM modes and the structural variations present in an ensemble (experimental/MD) are obtained with PCA. Similarity comparisons between a PC and a mode are evaluated by three measures defined by Tama and Sanejouand.[23]

***Overlap (O).*** This is a measure of how similar the direction of a given mode of motion $M_j$ from an ENM is in comparison with the PC eigenvector $P_i$ and is calculated as

$$O_{ij} = \frac{|P_i \cdot M_j|}{\|P_i\| \|M_j\|} \tag{11}$$

where $|P_i . M_j|$ refers to the absolute value of the dot product of $P_i$ and $M_j$ and $\|P_i\|$ and $\|M_j\|$ refer to the length of the PC and mode vectors, respectively. The sign of the dot product is not considered since the modes are harmonic in nature. The maximum overlap between any of the first $k$ modes of motion with the PC eigenvector $P_i$ is obtained as

$$O_i^{max} = \max_{j=1 \; to \; k} O_{ij} \tag{12}$$

***Cumulative Overlap (CO).*** This is a measure of how well a set of the first $k$ modes from an ENM capture the motion sampled by a single PC eigenvector $P_i$ and is calculated as

$$CO_i^k = \sqrt{\sum_{j=1}^{k} O_{ij}^2} \tag{13}$$

***Root Mean Square Inner Product (RMSIP).*** This quantity measures the similarity in directions between the set of first $k$ modes from an ENM and the first $l$ PC eigenvectors from a structural ensemble as

$$RMSIP_l^k = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{k}\left(P_i \cdot M_j\right)} \tag{14}$$

Based on the above three measures, we use ten different performance metrics to evaluate the performance of elastic network models in comparison to PCs from an ensemble as follows: the maximum overlap between the first 20 modes from the ENM and each of PC1 ($O_1^{max}$), PC2 ($O_2^{max}$) and PC3 ($O_3^{max}$); the cumulative overlap between the first 20 modes from the ENM and PC1 ($CO_1^{20}$), PC2 ($CO_2^{20}$) and PC3 ($CO_3^{20}$); and the RMSIP between the first 20 ANM modes and sets of the first 3 ($RMSIP_3^{20}$), 6 ($RMSIP_6^{20}$), 10 ($RMSIP_{10}^{20}$) and 20 PCs ($RMSIP_{20}^{20}$).

In addition, Pearson's correlation coefficient is reported between the calculated B-factors ($B^{calc}$) from the ENM and the crystallographic temperature factors ($B^{exp}$) from the representative structure in the experimental ensemble as

$$\rho^{exp,calc} = \frac{\boldsymbol{B}^{exp}-\langle\boldsymbol{B}^{exp}\rangle}{\|\boldsymbol{B}^{exp}-\langle\boldsymbol{B}^{exp}\rangle\|}\,\frac{\boldsymbol{B}^{calc}-\langle\boldsymbol{B}^{calc}\rangle}{\|\boldsymbol{B}^{calc}-\langle\boldsymbol{B}^{calc}\rangle\|} \tag{15}$$

## RESULTS AND DISCUSSION

**Comparison of ENM modes with the motions present within experimental structure ensembles.** We have previously shown for HIV-1 protease that the modes of motion from the classical ANM of a single structure correspond closely to the motions extracted from a set of experimental structures.[25] Several other studies have also demonstrated the power of ANMs in capturing the structural variations within experimental ensembles for a variety of proteins.[26,27] Here, we compare the motions predicted by the classical ANM and four other variants of ENMs with the motions present in experimental structure ensembles for a much larger dataset of 50 different proteins[40] (see Table S1).

In addition to the classical ANM, we use the four other types of modified ENMs (refer to Methods above for more details): (1) pfANM[34] with the spring constants between every residue pair weighted by the inverse of the sixth power of the distance between them; (2) eANM,[32] where the spring constants are weighted by the absolute values of the MJ potential energies between amino acid pair; (3) ccANM, in which the spring constants are weighted by the inverse of the contact-change based entropy value for each amino acid pair (based on our previous work);[39] and (4) dcANM[37] with the spring constant between every pair of residues weighted by the inverse of the variance of the internal distances between them (over all the structures in the experimental ensemble). The performance of each ANM is evaluated for the ten different metrics described in Methods.

We compute the motions for the ENMs of the representative structure from each protein ensemble (identified as the structure having the lowest RMSD from all other structures). Table 1 shows the average values (over the 50 proteins) of the 10 metrics for each type of ENM investigated. As expected, the dcANM naturally outperforms all of the other kinds of ANM in almost all the metrics. This is because the springs of the dcANM have been chosen directly from the internal distance changes between every pair of residues within the dataset for each protein; and hence it is naturally able to better reproduce the structural variations present in the dataset since it is built directly on the data being compared. The performance assessment of the other

ENMs against one another is more relevant to understanding the behavior of the ENMs. Based on the number of metrics for which the ENM is best, the ranking of the models is as follows: pfANM > ccANM > ANM > eANM. It is clear from Table 1 that the pfANM outperforms the other types of ENMs. Also, the ccANM performs essentially at the same level as the ANM on all 10 metrics.

**Table 1. Performance of different types of ENMs for the dataset of 50 proteins in comparison with the motions present in the experimental ensembles.**

| **Model** | $\boldsymbol{O_1^{max}}$ | $\boldsymbol{O_2^{max}}$ | $\boldsymbol{O_3^{max}}$ | $\boldsymbol{CO_1^{20}}$ | $\boldsymbol{CO_2^{20}}$ | $\boldsymbol{CO_3^{20}}$ | $\boldsymbol{RMSIP_3^{20}}$ | $\boldsymbol{RMSIP_6^{20}}$ | $\boldsymbol{RMSIP_{10}^{20}}$ | $\boldsymbol{RMSIP_{20}^{20}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| ANM | **0.41**<br>± 0.20 | 0.42<br>± 0.18 | 0.44<br>± 0.14 | 0.67<br>± 0.21 | 0.70<br>± 0.18 | 0.71<br>± 0.13 | 0.70<br>± 0.13 | 0.67<br>± 0.09 | 0.63<br>± 0.07 | 0.56<br>± 0.06 |
| pfANM | 0.39<br>± 0.19 | **0.44**<br>± 0.17 | **0.45**<br>± 0.13 | **0.68**<br>± 0.21 | **0.73**<br>± 0.17 | **0.74**<br>± 0.12 | **0.73**<br>± 0.12 | **0.70**<br>± 0.09 | **0.66**<br>± 0.07 | **0.59**<br>± 0.06 |
| eANM | 0.39<br>± 0.20 | 0.42<br>± 0.18 | 0.44<br>± 0.14 | 0.66<br>± 0.22 | 0.70<br>± 0.18 | 0.72<br>± 0.13 | 0.70<br>± 0.13 | 0.66<br>± 0.10 | 0.63<br>± 0.08 | 0.56<br>± 0.06 |
| ccANM | **0.41**<br>± 0.20 | 0.43<br>± 0.18 | 0.44<br>± 0.13 | 0.67<br>± 0.22 | 0.71<br>± 0.17 | 0.72<br>± 0.12 | 0.71<br>± 0.13 | 0.67<br>± 0.10 | 0.64<br>± 0.07 | 0.57<br>± 0.06 |
| dcANM* | *0.56*<br>*± 0.19* | *0.49*<br>*± 0.15* | *0.50*<br>*± 0.13* | *0.83*<br>*± 0.14* | *0.82*<br>*± 0.12* | *0.82*<br>*± 0.10* | *0.83*<br>*± 0.08* | *0.78*<br>*± 0.07* | *0.73*<br>*± 0.06* | *0.64*<br>*± 0.06* |

Values for each metric (as defined in Methods) are averaged over the 50 proteins. Values for the best performing model for each metric are shown in bold. Standard deviations are given as ± values
*dcANM is trained using the variances of the internal distance changes between residues in each experimental ensemble, and results are shown in italics.

**Comparison with protein motions from MD and experimental datasets.** Often only one structure of a protein or its close homolog is available. In such cases, a conformational sampling of the protein is often obtained using various computational techniques such as MD or Monte Carlo simulations. Once the simulation is run, the set of resulting structures are aligned to the starting structure and the 'essential motions'[5] extracted from the trajectory using PCA as described in the Methods section.

We performed a sequence-based search on the MODEL database[43], an online repository of MD simulations for available MD trajectories of the proteins or their homologs present in the dataset of 50 proteins. We identify 17 proteins for which MD simulation data were available for the protein or a substantial part of it (Table S2). We then compare how well the motions sampled by MD simulations for the set of 17 proteins compare against the variations present in sets of experimental structures of the same protein. Table 2 shows this comparison of the PCs extracted from the experimental dataset vs MD dataset for the 17 proteins.

**Table 2. Comparison of MD and experimental motions for the set of 17 proteins.**

| **Metric** | $\boldsymbol{O_1^{max}}$ | $\boldsymbol{O_2^{max}}$ | $\boldsymbol{O_3^{max}}$ | $\boldsymbol{CO_1^{20}}$ | $\boldsymbol{CO_2^{20}}$ | $\boldsymbol{CO_3^{20}}$ | $\boldsymbol{RMSIP_3^{20}}$ | $\boldsymbol{RMSIP_6^{20}}$ | $\boldsymbol{RMSIP_{10}^{20}}$ | $\boldsymbol{RMSIP_{20}^{20}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Value | 0.37<br>± 0.18 | 0.37<br>± 0.13 | 0.37<br>± 0.10 | 0.63<br>± 0.18 | 0.65<br>± 0.17 | 0.70<br>± 0.10 | 0.67<br>± 0.12 | 0.65<br>± 0.09 | 0.62<br>± 0.07 | 0.56<br>± 0.05 |
| Values for each metric (as defined in Methods) are averaged over the 17 proteins. | | | | | | | | | | |

The average maximum overlap between the first twenty PC directions from the MD ensemble with the PC1, PC2 and PC3 of the experimental ensemble is 0.37; which is comparatively smaller than the average values obtained for the classical ANM or any of the variants of the ANM. This difference is small and thus probably not significant. Several factors can affect the set of structures sampled in the MD trajectory; including the force field used, the simulation time, etc. It is also possible that the overlap between the conformational space sampled by MD and experiments is relatively small. As a result, a dcANM trained on the MD dataset could not reproduce well the set of motions in the experimental (MD) ensemble (Table S3). The fact that the ENMs reproduce the experimental ensemble better is noteworthy.

In order to further demonstrate that the motions sampled by MD and the experimental ensembles are often different, we provide two examples of dynamical cross-correlations (DCCMs)[45] of the residues from experimental and MD datasets for two different proteins in the dataset, lysozyme C (Figure 1A and B) and human leukocyte antigen (HLA) class II histocompatibility antigen alpha chain (HLA-DRA) (Figure 1C and D). These were chosen to demonstrate outliers in terms of being most similar and most different. In the case of lysozyme C, the two DCCMs are similar but with intricate differences, whereas in the case of the HLA-DRA, there is major differences between the correlations shown.

A closer inspection of the plots for HLA-DRA reveals that in the MD dataset, there are stronger correlations among the residues within each of its two domains ($\alpha 1$ and $\alpha 2$), particularly for $\alpha 2$, suggesting that the domains move almost as if they were rigid bodies. On the other hand, within the experimental ensemble, the higher correlations mostly correspond to residues within the same secondary structure, which can be easily identified from the plots. In other words, higher variabilities are observed in the relative orientations of the secondary structures within each domain. Previous studies have also shown that the DCCMs of the same protein from distinct simulations over different time-scales in MD simulations can be different.[46] Our results further support these observations in addition to suggesting that the dynamical cross correlations observed in MD often do not correspond to those observed in a set of experimental structures.

S Since the lengths of the MD simulations differ, one possible reason for the low level of agreement between motions from experiments and the simulations is a short simulation time. In order to ascertain whether this is the case, we divide the dataset into two sets: short (< 80ns) and long (≥ 80 ns) simulations (see Table S4). We then perform hypothesis testing to see whether the average values for each of the ten metrics for the short simulations are worse than those for the long simulations. Our analysis suggests that the observed differences are not significant (the p-values for all metrics are > 0.4), at least for the current dataset (Table S4). More detailed studies on larger datasets would be needed to reach a more certain conclusion.

**Comparison of ENM modes with motions present in MD structural ensembles.** It is also interesting to test whether the motions predicted by ENMs correlate with the set of motions sampled by MD simulations of the same protein. Starting from the representative structure for each protein, we construct the different types of ENMs and investigate how well the modes compare with the motions extracted by PCA from the MD structural ensemble for each of the 17 proteins. Table 3 shows the average values for each of the ten performance metrics for the different types of ENMs.

Again, as expected the dcANM performs the best in all metrics reflecting the fact that it was trained on the dataset itself. The other different ENMs rank in the following order for the ten performance metrics: pfANM > ccANM > ANM > eANM. And, this is the same order as seen in Table 1. It can be seen that the pfANM systematically outperforms the other types of ANM in reproducing the protein motions in the MD dataset, even though by a small margin. Taken together with the results from the performance on the experimental dataset, this seems to suggest that the overall intrinsic dynamics of the protein is dictated primarily by its geometry, i.e., the distances of separation between all pairs of different residues. The specific amino acid interactions of course allow the protein perform its specific functions; and will account for the differences in behaviors of various mutants of the protein; however, they do not much affect its global motions.

**Table 3. Performance of different types of ENMs in comparison with the motions present in the MD dataset of 17 proteins.**

| Model | $\boldsymbol{O_1^{max}}$ | $\boldsymbol{O_2^{max}}$ | $\boldsymbol{O_3^{max}}$ | $\boldsymbol{CO_1^{20}}$ | $\boldsymbol{CO_2^{20}}$ | $\boldsymbol{CO_3^{20}}$ | $\boldsymbol{RMSIP_3^{20}}$ | $\boldsymbol{RMSIP_6^{20}}$ | $\boldsymbol{RMSIP_{10}^{20}}$ | $\boldsymbol{RMSIP_{20}^{20}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| ANM | 0.42<br>± 0.14 | 0.46<br>± 0.22 | 0.40<br>± 0.12 | 0.68<br>± 0.17 | 0.68<br>± 0.20 | 0.68<br>± 0.14 | 0.68<br>± 0.15 | 0.70<br>± 0.11 | 0.70<br>± 0.07 | 0.65<br>± 0.05 |
| pfANM | **0.44**<br>± 0.14 | **0.46**<br>± 0.21 | **0.42**<br>± 0.09 | **0.71**<br>± 0.15 | **0.69**<br>± 0.20 | **0.71**<br>± 0.15 | **0.71**<br>± 0.15 | **0.73**<br>± 0.10 | **0.74**<br>± 0.06 | **0.70**<br>± 0.04 |
| eANM | 0.41<br>± 0.15 | 0.45<br>± 0.22 | 0.39<br>± 0.12 | 0.67<br>± 0.18 | 0.67<br>± 0.21 | 0.66<br>± 0.15 | 0.67<br>± 0.16 | 0.69<br>± 0.12 | 0.69<br>± 0.08 | 0.65<br>± 0.06 |
| ccANM | 0.43<br>± 0.14 | 0.44<br>± 0.21 | 0.40<br>± 0.11 | 0.70<br>± 0.16 | 0.68<br>± 0.20 | 0.68<br>± 0.14 | 0.69<br>± 0.15 | 0.71<br>± 0.10 | 0.71<br>± 0.07 | 0.66<br>± 0.05 |
| dcANM* | *0.51*<br>*± 0.15* | *0.44*<br>*± 0.13* | *0.44*<br>*± 0.13* | *0.80*<br>*± 0.16* | *0.77*<br>*± 0.16* | *0.75*<br>*± 0.16* | *0.78*<br>*± 0.15* | *0.74*<br>*± 0.10* | *0.71*<br>*± 0.07* | *0.63*<br>*± 0.05* |

Values for each metric (as defined in Methods) are averaged over the 17 proteins. Values for the best performing model are shown in bold and the next best in italics.
*dcANM is trained using the variances of the internal distance changes between residues in each MD ensemble, and results are shown in italics.

A comparison between the results in Table 1 and Table 3 shows a remarkable similarity in the abilities of the various ENMs to reproduce the motions in the ensembles of both the experimental sets of structures and the MD ensembles.

**Performance of ENMs in reproducing crystallographic B-factors.** In addition to being able to reproduce intrinsic protein motions, another strength of the ENMs is in their being able to reproduce crystallographic temperature factors of the residues in the protein. Here we generate different types of ENMs using the representative structure for each of the 17 proteins with MD trajectory data and compute B-factors from the models (see Methods). The dcANM models are generated by adjusting the spring constants using the internal distance changes present in the experimental and MD ensembles as described before. We then compute the Pearson's correlations between the predicted B-factors and the crystallographic B-factors of the representative structure in the experimental ensemble (Table 4).

**Table 4. Correlation between experimental temperature factors and predicted B-factors from various types of ANMs on the experimental and MD datasets.**

| Model | Correlation (MD dataset)* | Correlation (experimental dataset)# |
|---|---|---|
| ANM | 0.50 ± 0.14 | 0.53 ± 0.14 |
| pfANM | 0.52 ± 0.17 | **0.56** ± 0.14 |
| eANM | 0.51 ± 0.13 | 0.53 ± 0.12 |
| ccANM | 0.48 ± 0.14 | 0.50 ± 0.14 |
| dcANM | **0.53** ± 0.20 | 0.51 ± 0.18 |

Values are averaged over the 17 proteins. Value for the best performing model is shown in bold.
*dcANM is trained using internal distance changes between residues in the MD dataset;
#dcANM is trained using internal distance changes between residues in the experimental dataset;
Correlation values are with the crystallographic B-factors of the experimental representative structure.

As can be seen in Table 4, the pfANM gives the highest correlation with crystallographic B-factors. The dcANM model based on the MD dataset gives only a slightly better correlation with B-factors than the pfANM and is probably not a significant difference. Our results also confirm the observation by Hamacher and McCammon[32] that the eANM provides slight gains over the ANM in its being able to predict crystallographic B-factors (at least for the cases in the MD dataset). However, the values in Table 4 are all very similar. Interestingly, the eANM is slightly worse than the classical ANM or the ccANM at predicting motions present in the experimental ensembles as seen above (Tables 1 and 3). On the other hand, it is slightly better than the ccANM at reproducing crystallographic B-factors. This is in close agreement with observations by Fuglebakk and others[47] that a higher correlation with B-factors usually comes at the expense of the ability to predict collective protein motions.

## CONCLUSIONS

In this study, we have systematically compared the motions extracted from experimental structure ensembles of 50 different proteins with the motions predicted using several different variants of ENMs. In addition to the classic ANM, we study several modified ANMs which account more specifically for the geometry of the protein (pfANM and dcANM) or for the amino acid specificity of the residues, either in energy (eANM) or in entropy (ccANM). The ccANM is a new model introduced in this paper, which accounts for the relative entropies of amino acid pairs; which were derived from the relative frequencies of contact changes within a set of experimental protein conformational changes.  Our results show that pfANMs (taking into account all distances between residues in a protein structure) are best in capturing the structural variations present within an experimental ensemble of the same protein. The ccANMs do perform better than eANMs and the classic ANMs suggesting that the pair-wise entropies are important for conformational changes. The main conclusion is that the distances of separation between residues (i.e. the geometry in pfANM) plays a larger role than the chemical nature of the interactions (as in eANM or ccANM) for the overall intrinsic dynamics of proteins. Interestingly this is consistent with the strong dependence on geometry (shape) for the slowest motions,[48,49] supporting the overall viewpoint implicit in the elastic network models that geometry alone is important for the important protein dynamics.

In addition, we also have collected large scale molecular dynamics simulation data available for 17 proteins in the dataset and compared their structural changes with the structural variations present in the experimental set and those predicted by different types of ANM. The correspondences observed between the MD and experimental datasets is relatively poor when compared to the ANMs, highlighting some of the possible sampling problems in MD datasets, such as the force-field used, and simulation times. We also observe that training ANMs based on internal distance changes between residues observed in an MD simulation (dcANM) does not necessarily improve the correspondence with experimental motions, at least for the dataset of 17 proteins investigated in this study.

We find that some ANMs, specifically the pfANM or ccANM give better agreement with experimental motions extracted from experimental or MD ensembles. On the other hand, they provide only relatively small improvements in terms of the correlation with experimental B-factors, in agreement with previous studies. However, as observed by others[47], we also find that agreement with B-factors and the ability to reproduce collective motions do not necessarily go together.

## ASSOCIATED CONTENT

**Supporting Information**.

The following supporting materials are available free of charge on the ACS Publications website at DOI: http://pubs.acs.org/?.

Table S1, List of representative structures in the protein experimental ensembles; Table S2, Protein MD trajectory data; Table S3, Performance of dcANM models on the experimental vs MD datasets.

## AUTHOR INFORMATION

### Corresponding Author

*Email: jernigan@iastate.edu.

### Author Contributions

KS and RLJ conceptualized the research; KS and SKM performed the research; KS and RLJ wrote the paper with help from SKM. All authors have given approval to the final version of the manuscript.

### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENT

This work was supported by National Institutes of Health Grant R01-GM72014 and National Science Foundation Grant MCB-1021785, as well as funds from the Carver Trust awarded to the Roy J. Carver Department of Biochemistry, Biophysics, and Molecular Biology. KS was also supported by fellowship funds from the Office of Biotechnology, Iowa State University. The authors thank all members of the Jernigan Lab and the Guang Song Lab at Iowa State University for frequent helpful discussions.

FIGURE

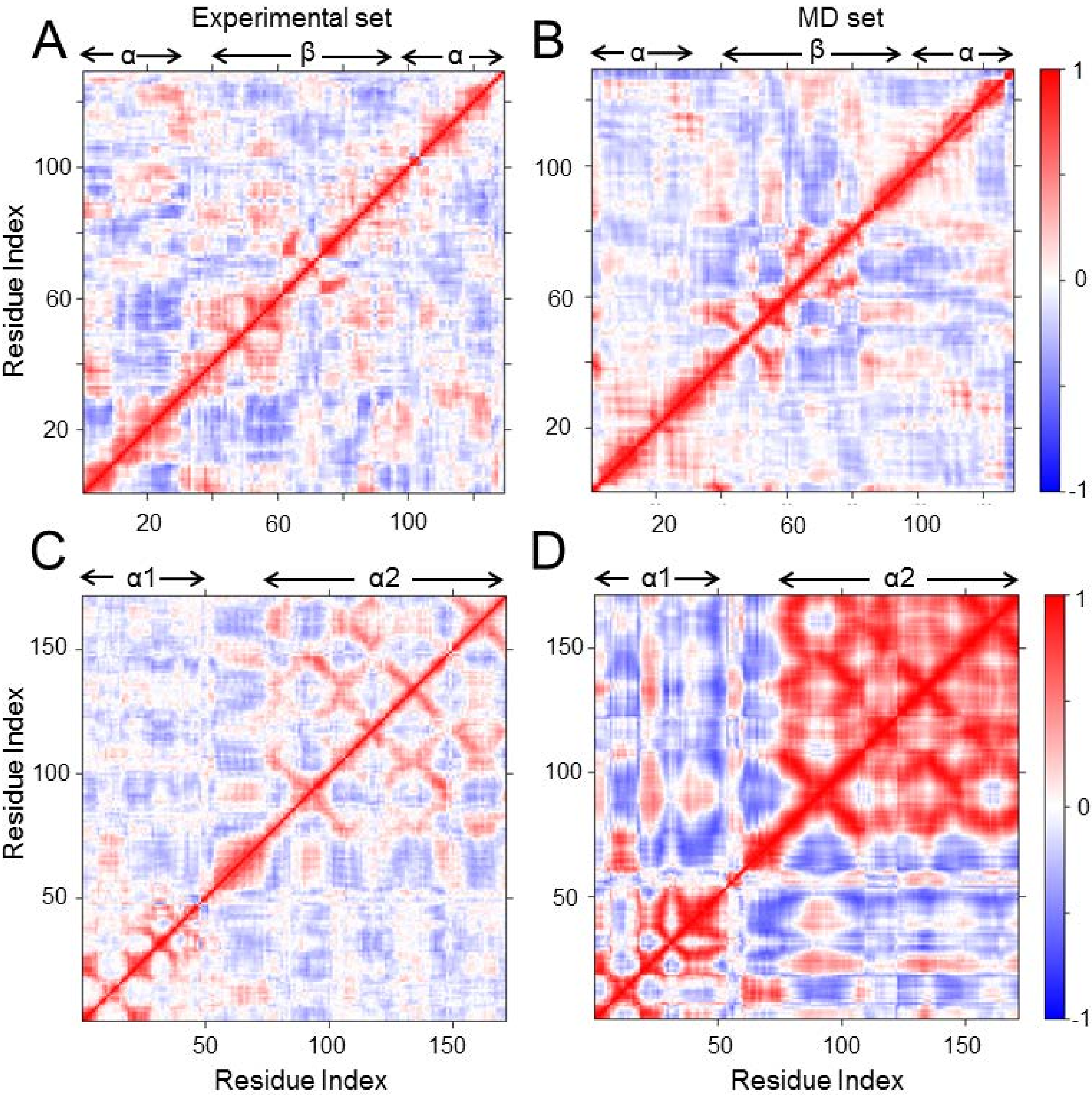

**Figure 1. Comparison of dynamical cross-correlation matrices (DCCMs) between experimental and MD datasets for lysozyme C (A, B) and HLA-DRA (C, D).** Positive correlations between residues are shown red and negative correlations in blue. The two domains (α and β) of lysozyme C and HLA-DR (α1 and α2) are indicated on top of the respective plots.

## REFERENCES

(1) Kohn, J. E.; Afonine, P. V; Ruscio, J. Z.; Adams, P. D.; Head-Gordon, T. Evidence of Functional Protein Dynamics from X-Ray Crystallographic Ensembles. *PLoS Comput. Biol.* **2010**, *6* (8), 1–5.

(2) Fenwick, R. B.; van den Bedem, H.; Fraser, J. S.; Wright, P. E. Integrated Description of Protein Dynamics from Room-Temperature X-Ray Crystallography and NMR. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (4), E445-54.

(3) Fernandez-Leiro, R.; Scheres, S. H. W. Unravelling Biological Macromolecules with Cryo-Electron Microscopy. *Nature* **2016**, *537* (7620), 339–346.

(4) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.* **1901**, *2* (11), 559–572.

(5) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins* **1993**, *17*, 412–425.

(6) Amadei, A.; Linssen, A. B.; de Groot, B. L.; van Aalten, D. M.; Berendsen, H. J. An Efficient Method for Sampling the Essential Subspace of Proteins. *J. Biomol. Struct. Dyn.* **1996**, *13* (4), 615–625.

(7) van Aalten, D. M.; Conn, D. A.; de Groot, B. L.; Berendsen, H. J.; Findlay, J. B.; Amadei, A. Protein Dynamics Derived from Clusters of Crystal Structures. *Biophys. J.* **1997**, *73* (6), 2891–2896.

(8) Howe, P. W. Principal Components Analysis of Protein Structure Ensembles Calculated Using NMR Data. *J. Biomol. NMR* **2001**, *20* (1), 61–70.

(9) Teodoro, M. L.; Phillips, G. N.; Kavraki, L. E. A Dimensionality Reduction Approach to Modeling Protein Flexibility. *Proc. sixth Annu. Int. Conf. Comput. Biol. RECOMB 02* **2002**, 299–308.

(10) Teodoro, M. L.; Phillips Jr., G. N.; Kavraki, L. E. Understanding Protein Flexibility through Dimensionality Reduction. *J. Comput. Biol.* **2003**, *10*, 617–634.

(11) Sanejouand, Y.-H. Elastic Network Models: Theoretical and Empirical Foundations. In *Biomolecular Simulations. Methods in Molecular Biology (Methods and Protocols)*; Monticelli, L., Salonen, E., Eds.; Humana Press: New York, 2013; Vol. 924, pp 601–616.

(12) Jernigan, R. L.; Yang, L.; Song, G.; Kurkckuoglu, O.; Doruker, P. Elastic Network Models of Coarse-Grained Proteins Are Effective for Studying the Structural Control Exerted over Their Dynamics. In *Coarse-Graining of Condensed Phase and Biomolecular Systems*; Voth, G. A., Ed.; CRC Press: Boca Raton, FL, 2009; pp 237–254.

(13) Chennubhotla, C.; Rader, A. J.; Yang, L.-W.; Bahar, I. Elastic Network Models for Understanding Biomolecular Machinery: From Enzymes to Supramolecular Assemblies. *Phys. Biol.* **2005**, *2* (4), S173–S180.

(14) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chem. Rev.* **2010**, *110* (3), 1463–1497.

(15) Setny, P.; Zacharias, M. Elastic Network Models of Nucleic Acids Flexibility. *J. Chem. Theory Comput.* **2013**, *9* (12), 5460–5470.

(16) Wang, Y.; Rader, A. J.; Bahar, I.; Jernigan, R. L. Global Ribosome Motions Revealed with Elastic Network Model. *J. Struct. Biol.* **2004**, *147* (3), 302–314.

(17) Wang, Y.; Jernigan, R. L. Comparison of tRNA Motions in the Free and Ribosomal Bound Structures. *Biophys. J.* **2005**, *89* (5), 3399–3409.

(18) Burton, B.; Zimmermann, M. T.; Jernigan, R. L.; Wang, Y. A Computational Investigation on the Connection between Dynamics Properties of Ribosomal Proteins and Ribosome Assembly. *PLoS Comput. Biol.* **2012**, *8* (5), e1002530.

(19) Yang, Z.; Májek, P.; Bahar, I. Allosteric Transitions of Supramolecular Systems Explored by Network Models: Application to Chaperonin GroEL. *PLoS Comput. Biol.* **2009**, *5* (4), e1000360.

(20) Keskin, O.; Bahar, I.; Flatow, D.; Covell, D. G.; Jernigan, R. L. Molecular Mechanisms of Chaperonin GroEL-GroES Function. *Biochemistry* **2002**, *41* (2), 491–501.

(21) Yang, L.; Song, G.; Jernigan, R. L. Comparisons of Experimental and Computed Protein Anisotropic Temperature Factors. *Proteins* **2009**, *76* (1), 164–175.

(22) Soheilifard, R.; Makarov, D. E.; Rodin, G. J. Critical Evaluation of Simple Network Models of Protein Dynamics and Their Comparison with Crystallographic B-Factors. *Phys. Biol.* **2008**, *5* (2), 026008.

(23) Tama, F.; Sanejouand, Y. H. Conformational Change of Proteins Arising from Normal

Mode Calculations. *Protein Eng.* **2001**, *14* (1), 1–6.

(24) Yang, L.; Song, G.; Jernigan, R. L. How Well Can We Understand Large-Scale Protein Motions Using Normal Modes of Elastic Network Models? *Biophys. J.* **2007**, *93* (3), 920–929.

(25) Yang, L.; Song, G.; Carriquiry, A.; Jernigan, R. L. Close Correspondence between the Motions from Principal Component Analysis of Multiple HIV-1 Protease Structures and Elastic Network Modes. *Structure* **2008**, *16* (2), 321–330.

(26) Yang, L.-W.; Eyal, E.; Bahar, I.; Kitao, A. Principal Component Analysis of Native Ensembles of Biomolecular Structures (PCA_NEST): Insights into Functional Dynamics. *Bioinformatics* **2009**, *25* (5), 606–614.

(27) Skjaerven, L.; Martinez, A.; Reuter, N. Principal Component and Normal Mode Analysis of Proteins; a Quantitative Comparison Using the GroEL Subunit. *Proteins Struct. Funct. Bioinforma.* **2011**, *79* (1), 232–243.

(28) Zimmermann, M. T.; Jernigan, R. L. Elastic Network Models Capture the Motions Apparent within Ensembles of RNA Structures. *RNA* **2014**, *20* (6), 792–804.

(29) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80* (1), 505–515.

(30) Kim, M. H.; Seo, S.; Jeong, J. Il; Kim, B. J.; Liu, W. K.; Lim, B. S.; Choi, J. B.; Kim, M. K. A Mass Weighted Chemical Elastic Network Model Elucidates Closed Form Domain

Motions in Proteins. *Protein Sci.* **2013**, *22* (5), 605–613.

(31) Frappier, V.; Najmanovich, R. J. A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. *PLoS Comput. Biol.* **2014**, *10* (4), e1003569.

(32) Hamacher, K.; McCammon, J. A. Computing the Amino Acid Specificity of Fluctuations in Biomolecular Systems. *J. Chem. Theory Comput.* **2006**, *2* (3), 873–878.

(33) Miyazawa, S.; Jernigan, R. L. Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* **1996**, *256* (3), 623–644.

(34) Yang, L.; Song, G.; Jernigan, R. L. Protein Elastic Network Models and the Ranges of Cooperativity. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (30), 12347–12352.

(35) Moritsugu, K.; Smith, J. C. Coarse-Grained Biomolecular Simulation with REACH: Realistic Extension Algorithm via Covariance Hessian. *Biophys. J.* **2007**, *93* (10), 3460–3469.

(36) Lyman, E.; Pfaendtner, J.; Voth, G. A. Systematic Multiscale Parameterization of Heterogeneous Elastic Network Models of Proteins. *Biophys. J.* **2008**, *95* (9), 4183–4192.

(37) Katebi, A. R.; Sankar, K.; Jia, K.; Jernigan, R. L. The Use of Experimental Structures to Model Protein Dynamics. In *Molecular Modeling of Proteins*; 2015; Vol. 1215, pp 213–236.

(38) Skjærven, L.; Yao, X.-Q.; Scarabelli, G.; Grant, B. J. Integrating Protein Structural

Dynamics and Evolutionary Analysis with Bio3D. *BMC Bioinformatics* **2014**, *15* (1), 399.

(39) Sankar, K.; Jia, K.; Jernigan, R. L. Knowledge-Based Entropies Improve the Identification of Native Protein Structures. *Proc. Natl. Acad. Sci.* **2017**, *114* (11), 2928–2933.

(40) Sankar, K.; Liu, J.; Wang, Y.; Jernigan, R. L. Distributions of Experimental Protein Structures on Coarse-Grained Free Energy Landscapes. *J. Chem. Phys.* **2015**, *143* (24), 243153.

(41) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(42) Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. MUSTANG: A Multiple Structural Alignment Algorithm. *Proteins* **2006**, *64* (3), 559–574.

(43) Meyer, T.; D’Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Pérez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; et al. MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure* **2010**, *18*, 1399–1409.

(44) Sievers, F.; Higgins, D. G. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. *Methods Mol. Biol.* **2014**, *1079*, 105–116.

(45) Ichiye, T.; Karplus, M. Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins* **1991**, *11* (3), 205–217.

(46) Hünenberger, P. H.; Mark, A. E.; van Gunsteren, W. F. Fluctuation and Cross-Correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J. Mol. Biol.* **1995**, *252* (4), 492–503.

(47) Fuglebakk, E.; Reuter, N.; Hinsen, K. Evaluation of Protein Elastic Network Models Based on an Analysis of Collective Motions. *J. Chem. Theory Comput.* **2013**, *9* (12), 5618–5628.

(48) Doruker, P.; Jernigan, R. L. Functional Motions Can Be Extracted from on-Lattice Construction of Protein Structures. *Proteins Struct. Funct. Genet.* **2003**, *53* (2), 174–181.

(49) Ma, J. New Advances in Normal Mode Analysis of Supermolecular Complexes and Applications to Structural Refinement. *Curr. Protein Pept. Sci.* **2004**, *5* (2), 119–123.

# Supporting Information

# for

# Comparisons of Protein Dynamics from Experimental Structure Ensembles, Molecular Dynamics Ensembles, and Coarse-Grained Elastic Network Models

*Kannan Sankar* [†‡§], *Sambit K. Mishra*[†‡] *and Robert L. Jernigan*[†‡*]

[†]Bioinformatics and Computational Biology Interdepartmental Graduate Program, Iowa State University, Ames, IA 50011-1178, USA

[‡]Roy J. Carver Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA 50011-1178, USA

[*]Corresponding author: E-mail: jernigan@iastate.edu

## List of Supplementary Information Files

**Table S1**. **List of proteins in the experimental ensemble dataset.**

| Set # | Protein Name | #Residues | #Structures | Organism | Representative Structure |
|---|---|---|---|---|---|
| 1 | Sarcoplasmic/endoplasmic reticulum calcium ATPase 1 isoform (SERCA1a) | 995 | 63 | *Oryctolagus cuniculus* | 3NAL_A |
| 2 | Peptidyl-prolyl cis-trans isomerase A | 159 | 136 | *Homo sapiens* | 3ODL_A |

| | | | | | |
|---|---|---|---|---|---|
| 3 | Human Lysozyme C | 131 | 218 | *Homo sapiens* | 1B5U_A |
| 4 | B. anthracis Dihydrofolate reductase (DHFR) | 162 | 76 | *Bacillus anthracis* | 3FL8_F |
| 5 | Cytochrome c peroxidase, mitochondrial | 292 | 165 | *Saccharomyces* | 2AQD_A |
| 6 | HLA class II histocompatibility antigen, D-R alpha | 172 | 108 | *Homo sapiens* | 1T5W_A |
| 7 | Thaumatin I | 202 | 80 | *Thaumatococcus* | 3AOK_A |
| 8 | FK506-binding protein | 108 | 59 | *Homo sapiens* | 1D6O_A |
| 9 | Human serum albumin (HSA) | 555 | 99 | *Homo sapiens* | 2BXB_B |
| 10 | Phi6 RNA-directed RNA polymerase | 665 | 55 | *Pseudomonas phage* | 1UVJ_A |
| 11 | Squalene synthase | 332 | 61 | *Homo sapiens* | 3WCF_F |
| 12 | Camphor 5-monooxygenase | 402 | 134 | *Pseudomonas* | 1UYU_B |
| 13 | Azurin | 129 | 202 | *Pseudomonas* | 1E5Y_C |
| 14 | Proteinase K | 280 | 61 | *Engyodontium* | 3DVR_X |
| 15 | Beta-lactamase | 359 | 143 | *Escherichia coli* | 4KZ5_B |
| 16 | Hepatitis C RNA-directed RNA polymerase | 548 | 162 | *Hepatitis C virus* | 2XHU_B |
| 17 | Tankyrase-2 | 186 | 64 | *Homo sapiens* | 4PNN_B |
| 18 | Heparin-binding growth factor 1 | 122 | 61 | *Homo sapiens* | 2HW9_A |
| 19 | Casein kinase II subunit alpha | 326 | 78 | *Homo sapiens* | 3NGA_A |
| 20 | Thioredoxin 1 | 104 | 80 | *Escherichia coli* | 2H73_A |
| 21 | H-2 class I histocompatibility antigen, alpha chain | 272 | 89 | *Mus musculus* | 1S7U_A |
| 22 | T4 lysozyme | 163 | 183 | *Enterobacteria* | 1G0J_A |
| 23 | GTPase HRas | 165 | 100 | *Homo sapiens* | 4L9W_A |
| 24 | Heparin-binding growth factor 1 | 121 | 130 | *Homo sapiens* | 1JQZ_A |
| 25 | Aldose reductase | 309 | 120 | *Homo sapiens* | 2IKH_A |
| 26 | Phosphopentomutase | 390 | 60 | *Bacillus cereus* | 3M8Z_B |
| 27 | MHC class I antigen | 274 | 64 | *Homo sapiens* | 1ZSD_A |
| 28 | Carboxypeptidase B | 304 | 58 | *Sus scrofa* | 2PJ5_B |
| 29 | HLA class I histocompatibility antigen, A-2 alpha | 276 | 256 | *Homo sapiens* | 3KLA_A |
| 30 | Chemotaxis protein CheY | 115 | 109 | *Escherichia coli* | 3F7N_B |
| 31 | DNA polymerase beta | 326 | 154 | *Homo sapiens* | 8ICZ_A |
| 32 | Human Dihydrofolate reductase | 183 | 74 | *Homo sapiens* | 1BOZ_A |
| 33 | Glucosylceramidase | 488 | 64 | *Homo sapiens* | 1OGS_B |
| 34 | D-alanyl-D-alanine Carboxypeptidase | 461 | 72 | *Actinomadura sp.* | 4BEN_C |
| 35 | WD repeat-containing protein 5 | 294 | 80 | *Homo sapiens* | 2H6Q_B |
| 36 | LeuT Transporter | 503 | 45 | *Aquifex aeolicus* | 3F3D_A |
| 37 | Cathepsin S | 217 | 58 | *Homo sapiens* | 2FRA_B |
| 38 | Thermolysin | 317 | 122 | *Bacillus* | 1KEI_A |
| 39 | Polymerase | 458 | 55 | *Human poliovirus 1* | 3OL6_A |
| 40 | Hen egg white lysozyme | 130 | 586 | *Gallus gallus* | 194L_A |
| 41 | Beta-2-microglobulin | 100 | 242 | *Mus musculus* | 1RJY_E |
| 42 | Phospholipase A2 | 122 | 80 | *Daboia russellii* | 1SV9_A |
| 43 | Beta-lactamase TEM | 260 | 59 | *Escherichia coli* | 1NYY_A |
| 44 | Guanyl-specific ribonuclease T1 | 105 | 89 | *Aspergillus oryzae* | 1BU4_A |
| 45 | E-coli Dihydrofolate reductase | 160 | 80 | *Escherichia coli* | 1DHI_B |
| 46 | Insulin-degrading enzyme | 942 | 61 | *Homo sapiens* | 3OFI_A |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 47 | Cationic trypsin | 224 | 421 | *Bos taurus* | 1S0Q_A |
| 48 | Elastase 1 | 241 | 116 | *Sus scrofa* | 2BD3_A |
| 49 | Endothiapepsin | 331 | 52 | *Endothia parasitica* | 3PI0_A |
| 50 | Macrophage metalloelastase | 153 | 83 | *Homo sapiens* | 3F17_A |

## Table S2. List of proteins with MD trajectory data.

| Set # | Protein Name | Organism | Representative PDB with MD data | Simulation Details |
|---|---|---|---|---|
| 1 | Beta-2-microglobulin | *Mus musculus* | 1HSA | Amber 8, 20 ns |
| 2 | Camphor 5-monooxygenase | *Pseudomonas putida* | 1AKD | Amber 8, 10.5 |
| 3 | H-2 class I histocompatibility antigen, D-B | *Mus musculus* | 1HSA | Amber 8, 20 ns |
| 4 | Thermolysin | *Bacillus* | 1FJ3 | Amber 8 v1, 10 |
| 5 | Cytochrome c peroxidase, mitochondrial | *Saccharomyces* | 1JDR | Amber 8, 10 ns |
| 6 | HLA class I histocompatibility antigen, A-2 | *Homo sapiens* | 2BVO | Amber 8, 20 ns |
| 7 | MHC class I antigen | *Homo sapiens* | 2AXG | Amber 8, 10 ns |
| 8 | Elastase 1 | *Sus scorfa* | 1ESA | Amber 9, 80 ns |
| 9 | Thaumatin I | *Thaumatococcus* | 1THV | Amber 9, 80 ns |
| 10 | HLA class II histocompatibility antigen, DR | *Homo sapiens* | 1DLH | Amber 8, 10 ns |
| 11 | Peptidyl-prolyl cis-trans isomerase A | *Homo sapiens* | 2CPL | Amber 8, 80.5 |
| 12 | Heparin-binding growth factor 1 | *Homo sapiens* | 1FMM | Amber 8, 10 ns |
| 13 | Hen Egg White Lysozyme C | *Gus gallus* | 1DPX | Amber 8, 20 ns |
| 14 | Heparin-binding growth factor 1 | *Gallus gallus* | 1FMM | Amber 8, 10 ns |
| 15 | Human Lysozyme C | *Homo sapiens* | 1JSF | Amber 8, 10 ns |
| 16 | Phospholipase A2 | *Daboia russellii* | 1BBC | Amber 8, 10 ns |
| 17 | FK506-binding protein | *Homo sapiens* | 1FKB | Amber 8v1, |

**Table S3**. **Comparison of dcANMs based on experimental and MD datasets.**

| Test<br>Train | $O_1^{max}$ | $O_2^{max}$ | $O_3^{max}$ | $CO_1^{20}$ | $CO_2^{20}$ | $CO_3^{20}$ | $RMSIP_3^{2}$ | $RMSIP_6^{2}$ | $RMSIP_1^{2}$ | $RMSIP_2^{2}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Exp<br>MD | 0.32 | 0.33 | 0.37 | 0.56 | 0.57 | 0.64 | 0.60 | 0.57 | 0.54 | 0.49 |
| MD<br>Exp | 0.34 | 0.32 | 0.30 | 0.60 | 0.56 | 0.55 | 0.58 | 0.58 | 0.56 | 0.50 |

Values for each metric are averaged over the 17 proteins.

The ‘Train’ set refers to the ensemble from which the internal distance changes were extracted to train the dcANM. The dcANM is built on the representative structure in each dataset. The ‘Test’ set refers to the ensemble from which the PCs were extracted. The modes from the dcANM generated using the ‘Train’ set are tested against the PCs from the ‘Test’ set using each of the 10 different metrics.

**Table S4**. **Comparison of performance metrics between short and long MD simulations**

| Representative PDB | Simulation Type | Simulation Time | $O_1^{max}$ | $O_2^{max}$ | $O_3^{max}$ | $CO_1^{20}$ | $CO_2^{20}$ | $CO_3^{20}$ | $RMSIP_3^2$ | $RMSIP_6^2$ | $RMSIP_1^2$ | $RMSIP_2^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1HSA | Short | 20 | 0.28 | 0.24 | 0.30 | 0.53 | 0.53 | 0.66 | 0.58 | 0.63 | 0.62 | 0.57 |
| 1AKD | Short | 10.5 | 0.38 | 0.31 | 0.25 | 0.73 | 0.61 | 0.62 | 0.66 | 0.61 | 0.58 | 0.53 |
| 1HSA | Short | 20 | 0.55 | 0.46 | 0.43 | 0.90 | 0.81 | 0.78 | 0.83 | 0.76 | 0.68 | 0.57 |
| 1FJ3 | Short | 10 | 0.55 | 0.32 | 0.43 | 0.78 | 0.58 | 0.82 | 0.73 | 0.67 | 0.59 | 0.51 |
| 1JDR | Short | 10 | 0.20 | 0.12 | 0.32 | 0.43 | 0.27 | 0.56 | 0.44 | 0.46 | 0.45 | 0.44 |
| 2BVO | Short | 20 | 0.76 | 0.62 | 0.43 | 0.89 | 0.91 | 0.71 | 0.84 | 0.78 | 0.70 | 0.62 |
| 2AXG | Short | 10 | 0.67 | 0.29 | 0.44 | 0.90 | 0.54 | 0.84 | 0.77 | 0.68 | 0.65 | 0.57 |
| 1ESA | Long | 80 | 0.22 | 0.36 | 0.23 | 0.48 | 0.63 | 0.65 | 0.59 | 0.62 | 0.62 | 0.55 |
| 1THV | Long | 80 | 0.17 | 0.41 | 0.53 | 0.38 | 0.66 | 0.82 | 0.64 | 0.60 | 0.59 | 0.52 |
| 1DLH | Short | 10 | 0.34 | 0.52 | 0.36 | 0.71 | 0.89 | 0.75 | 0.78 | 0.70 | 0.67 | 0.57 |
| 2CPL | Long | 80.5 | 0.13 | 0.21 | 0.21 | 0.31 | 0.42 | 0.54 | 0.43 | 0.52 | 0.58 | 0.56 |
| 1FMM | Short | 10 | 0.35 | 0.30 | 0.34 | 0.58 | 0.58 | 0.75 | 0.65 | 0.64 | 0.60 | 0.54 |
| 1DPX | Short | 20 | 0.37 | 0.33 | 0.31 | 0.56 | 0.76 | 0.62 | 0.65 | 0.70 | 0.70 | 0.67 |
| 1FMM | Short | 10 | 0.24 | 0.39 | 0.27 | 0.50 | 0.61 | 0.53 | 0.55 | 0.54 | 0.53 | 0.49 |
| 1JSF | Short | 10 | 0.47 | 0.56 | 0.39 | 0.72 | 0.79 | 0.74 | 0.75 | 0.69 | 0.69 | 0.62 |
| 1BBC | Short | 10 | 0.38 | 0.31 | 0.52 | 0.78 | 0.64 | 0.77 | 0.73 | 0.67 | 0.60 | 0.53 |
| 1FKB | Long | 100 | 0.32 | 0.54 | 0.46 | 0.58 | 0.84 | 0.73 | 0.73 | 0.74 | 0.70 | 0.62 |
| | | | | | | | | | | | | |
| P-value (Wilcoxon Test)* | | | 1.00 | 0.39 | 0.56 | 0.99 | 0.48 | 0.69 | 0.95 | 0.85 | 0.61 | 0.56 |
| P-value (Welch's t-test)# | | | 1.00 | 0.44 | 0.55 | 0.99 | 0.57 | 0.62 | 0.88 | 0.74 | 0.46 | 0.39 |

*Wilcoxon rank sum test with $H_o: \mu_S = \mu_L$ and with $H_A: \mu_S < \mu_L$

#Welch's t- test with $H_o: \mu_S = \mu_L$ and with $H_A: \mu_S < \mu_L$

(S = short simulations < 80 ns; L = long simulations ≥80 ns)