Entropy **2012**, *14*, 213-232; doi:10.3390/e14020213

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Quantitative Comparison of Conformational Ensembles

Kevin C. Wolfe and Gregory S. Chirikjian *

Department of Mechanical Engineering, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

* Author to whom correspondence should be addressed; E-Mail: gregc@jhu.edu.

Received: 1 January 2012; in revised form: 27 January 2012 / Accepted: 3 February 2012 / Published: 10 February 2012

Abstract: A number of measures have been used in the structural biology literature to compare the shapes or conformations of biological macromolecules. However, the issue of how to compare two ensembles of conformations has received far less attention. Herein, the problem of how to quantitatively compare two such ensembles is addressed in several different ways using concepts from probability and information theory. Ultimately, such metrics could be used in the evaluation of structure-prediction algorithms and the analysis of how conformational mobility is inhibited by bound ligands.

Keywords: loop entropy; Kullback-Leibler divergence; Gaussian distributions; elastic networks

1. Introduction

The comparison of conformations of biological macromolecules has traditionally been purely geometric and primarily concerned with measuring the difference of two fixed structures [1–6]. In contrast, the issue of comparing ensembles of conformations has received relatively little attention. This issue can be particularly important if the macromolecules in question have significant flexibility. For example, consider a protein that is comprised of a large relatively rigid body and a flexible loop. The shape of the loop may vary significantly over time while the shape of the rigid portion remains largely fixed. Given two conformations of such a protein, traditional geometric methods of comparison may weight the importance of the shape of the loop and the rigid body equally, whereas comparing two conformational ensembles of the same protein may reduce the importance of the shape of the loop based on the uncertainty of its location with respect to the rigid body. This difference is the primary focus of this paper.

Several groups have looked at distributions and ensembles of protein conformations from a variety of perspectives [7–10]. Additionally, the flexibility and associated entropy of proteins has been studied [11–18]. This includes the work by Damm and Carlson presented in [19]. They proposed a geometric evaluation of protein prediction based on Gaussian-weighted root mean square deviation (RMSD). However, their method differs significantly from what is presented here as their weights are based on the distance between atoms in two different conformations as opposed to the flexibility of the proteins individually.

The remainder of this section establishes terminology and definitions that are used throughout the paper, and reviews methods for quantitatively comparing two conformations or shapes of the same macromolecule. Such metrics are used both to compare structures in biological contexts and in evaluating the performance of algorithms in competitions such as CASP [20,21] and CAPRI [22,23]. The rest of the paper is then structured as follows. Section 2 describes how conformational ensembles can be described as Gaussian distributions on high-dimensional configuration spaces. Section 3 explores similarity measures between conformational ensembles. This includes information-theoretic measures based on the Maxwell-Boltzmann probability distribution and its marginals. Section 4 provides some numerical examples of the metrics explored relative to more traditional approaches. An appendix is included that provides some information for reducing the degrees of freedom for a coarse-grained elastic network.

1.1. Terminology and Definitions

In the literature, the words conformation and configuration are sometimes used interchangeably, and at other times are given slightly different meanings. Here *configuration* is used to mean a collection of particles that are arranged in space in a particular way. When, in addition, the constituents of a configuration are connected together in a specified way, the result is a *conformation*. Connectivity information is topological in nature, and so conformation = configuration + topology. For example, a conformation can be described by a set of coordinates $\{\mathbf{x}_i\}$ for $i = 1, ..., n$ and a connectivity matrix, Γ. Here $\mathbf{x}_i = [x_i, y_i, z_i]^T \in \mathbb{R}^3$ is a three-dimensional position vector, and $\Gamma = \Gamma^T \in \mathbb{R}^{n \times n}$ is a symmetric matrix consisting of 1's and 0's where 1 means that two points are connected and 0 means that they are not. The set of points by itself, without the connectivity information contained in Γ, would describe a configuration.

When the connections allow for some degree of mobility, the specific vectors in the set $\{\mathbf{x}_i\}$ can change, while Γ remains constant. The set of all possible conformations consistent with the constraints is described as an ensemble. In the case of a macromolecule, the set of allowable conformations is defined by constraints from the covalent structure and non-covalent interactions such as hydrogen bonds, steric repulsion, and hydrophobicity. Even when all of these constraints are imposed, a significant degree of conformational mobility exists.

In contrast to the concepts of configuration and conformation, the concept of shape is a coarser description. Shape is concerned with the external appearance rather than internal connectivity. Shape can be described in terms of the boundary of a conformation and, since it disregards Γ, contains even less information than a configuration. A description that lies somewhere between configuration and shape is that of *density*. For example, if $\rho_i(\mathbf{x})$ is the electron density of the i^{th} atomic nucleus in a structure containing n atoms, then the density of a whole macromolecule can be described as

$$\rho(\mathbf{x}) = \sum_{i=1}^{n} \rho_i(\mathbf{x} - \mathbf{x}_i)$$

For example, each $\rho_i(\mathbf{x})$ might be described as a concentrated Gaussian distribution. Estimates of shape can be obtained from density by evaluating level curves of the form $\rho(\mathbf{x}) = c$ where c is some threshold value.

1.2. Established Measures of Conformational and Shape Similarity

Two conformations $(\{\mathbf{x}_i\}, \Gamma)$ and $(\{\mathbf{x}'_i\}, \Gamma')$ are said to be *(geometrically) congruent* if $\Gamma = \Gamma'$ and it is possible to find a rigid-body transformation $g = (R, \mathbf{t})$ where $R \in SO(3)$ (the set of 3×3 rotation matrices) and $\mathbf{t} \in \mathbb{R}^3$ (the set of three-dimensional translation/position vectors) such that $\mathbf{x}'_i = R\mathbf{x}_i + \mathbf{t}$ for all values $i = 1, ..., n$. It is convenient to use the shorthand $g \cdot \mathbf{x}_i \doteq R\mathbf{x}_i + \mathbf{t}$. The set of all g is denoted as G, the group of rigid-body motions in three-dimensional space, and $g \cdot \mathbf{x}_i$ defines an “action” of G on $\mathbb{R}^n$.

Two configurations are said to be congruent if $\{\mathbf{x}'_i\} = \{g \cdot \mathbf{x}_i\}$. This is a weaker condition than the corresponding concept of congruence for conformation for two reasons: (1) there is no constraint on Γ, since there is no Γ as part of the definition of a configuration; (2) equality of sets does not require a correspondence between individual members of the set, and hence is a loser condition than when specific elements are equated. Two densities $\rho(\mathbf{x})$ and $\rho'(\mathbf{x})$ are said to be congruent if it is possible to find a $g \in G$ such that $\rho(\mathbf{x}) = \rho'(g \cdot \mathbf{x})$.

When two conformations, configurations, or shapes are not congruent, it is useful to assess how close they are to being congruent. These measures of “similarity” are based on some measure of distance [24]. For example,

$$d_p(\mathbf{x}, \mathbf{x}') \doteq \|\mathbf{x} - \mathbf{x}'\|_p = (|x - x'|^p + |y - y'|^p + |z - z'|^p)^{\frac{1}{p}} \quad (1)$$

can be used to measure the distance between points where $p = 1$ (the Manhattan metric) and $p = 2$ (Euclidean distance) are the most common values. Likewise, distance between two densities can be defined as

$$D_p(\rho, \rho') \doteq \left(\int_{\mathbb{R}^3} |\rho(\mathbf{x}) - \rho'(\mathbf{x})|^p \, d\mathbf{x} \right)^{\frac{1}{p}} \quad (2)$$

A common similarity measure between two conformations is

$$d_p(\{\mathbf{x}_i\}, \{\mathbf{x}'_i\}) \doteq \min_{g \in G} \left(\sum_{i=1}^{n} \frac{1}{n} [d_p(\mathbf{x}_i \, , \, g \cdot \mathbf{x}'_i)]^p \right)^{\frac{1}{p}} \quad (3)$$

When $p = 2$, this is the commonly used root-mean-square (RMSD) comparison. It is also possible to define differences in conformations without minimizing over G by establishing internal coordinates such as dihedral angles and bond lengths, and comparing these internal coordinates for each conformation, but this approach will not be pursued here.

In the case of two conformations (or configurations) consisting of identical particles without correspondence, the optimal-assignment metric [25,26]

$$\delta_p(\{\mathbf{x}_i\},\{\mathbf{x}'_i\}) \doteq \min_{g\in G}\min_{\pi\in\Pi_n}\left(\sum_{i=1}^{n}[d_p(\mathbf{x}_i\,,\,g\cdot\mathbf{x}'_{\pi(i)})]^p\right)^{\frac{1}{p}} \tag{4}$$

can be used where Π_n is the set of permutations that acts on the n indices that label the particle positions. In cases where subsets of particles are of the same type rather than all of them being the same, the minimization over Π_n in Equation (4) can be broken down into minimizations within each subset of similar particle types.

It is also possible to define similarity metrics of the form

$$d'_p(\{\mathbf{x}_i\},\{\mathbf{x}'_i\}) \doteq \min_{g\in G}\sum_{i=1}^{n} d_p(\mathbf{x}_i\,,\,g\cdot\mathbf{x}'_i)$$

and

$$\delta'_p(\{\mathbf{x}_i\},\{\mathbf{x}'_i\}) \doteq \min_{g\in G}\min_{\pi\in\Pi_n}\sum_{i=1}^{n} d_p(\mathbf{x}_i\,,\,g\cdot\mathbf{x}'_{\pi(i)})$$

but these are less commonly used because they require more effort to compute.

Shape similarity can be measured by evaluating

$$\Delta_p(\rho,\rho') \doteq \min_{g\in G} D_p(\rho, g\cdot\rho') \tag{5}$$

where [27]

$$g\cdot\rho'(\mathbf{x}) \doteq \rho'(g^{-1}\cdot\mathbf{x}) \tag{6}$$

It can be shown that all of these measures of similarity satisfy the formal properties of a metric.

If the densities are normalized so as to be probability density functions,

$$\int_{\mathbb{R}^3}\rho(\mathbf{x})\,d\mathbf{x} = \int_{\mathbb{R}^3}\rho'(\mathbf{x})\,d\mathbf{x} = 1$$

then the *Kullback-Leibler divergence* (or *relative entropy*) is defined as

$$D_{KL}(\rho\,\|\,\rho') \doteq \int_{\mathbb{R}^3}\rho(\mathbf{x})\log\left(\frac{\rho(\mathbf{x})}{\rho'(\mathbf{x})}\right)d\mathbf{x}$$

A similarity measure can be defined as

$$\Delta_{KL}(\rho\,\|\,\rho') \doteq \min_{g\in G} D_{KL}(\rho\,\|\,g\cdot\rho') \tag{7}$$

Unlike the similarity metrics discussed above, this one is not symmetric.

Whereas the similarity measures reviewed in this section are for individual conformations, configurations, or shapes, it can be useful to compare two sets (or ensembles) of such objects rather than two individual objects. The next section addresses continuous ensembles that can be described as probability densities in high-dimensional spaces.

2. Generating Conformational Ensembles

Consider a macromolecule with atomic nuclei in a single conformation given by $\{\mathbf{x}_i\}$, and let the mass of the i^{th} of these be m_i. Then the kinetic energy will be

$$T = \frac{1}{2}\sum_{i=1}^{n} m_i\, \dot{\mathbf{x}}_i \cdot \dot{\mathbf{x}}_i \;=\; \frac{1}{2}\dot{\mathbf{X}}^T[M]\dot{\mathbf{X}}$$

where $\dot{\mathbf{x}}$ denotes the derivative of $\mathbf{x}$ with respect to time and $\cdot$ denotes the dot product of vectors, $\mathbf{X} = [\mathbf{x}_1^T, \cdots, \mathbf{x}_n^T]^T \in \mathbb{R}^{3n}$, and $[M]$ is the $3n \times 3n$ diagonal matrix with blocks on the diagonal of the form $m_i\mathbb{I}_3$ for $i = 1, ..., n$.

The i^{th} conjugate momentum vector is

$$\mathbf{p}_i = \frac{\partial T}{\partial \dot{\mathbf{x}}_i} = m_i \dot{\mathbf{x}}_i$$

Concatenating these gives $\mathbf{P} = [\mathbf{p}_1^T, ..., \mathbf{p}_n^T]^T$, and so the kinetic energy can be written as

$$T \;=\; \frac{1}{2}\mathbf{P}^T[M]^{-1}\mathbf{P}$$

The potential energy for the macromolecule will be of the form $V = V(\mathbf{X}\,;\,\mathbf{X}_0)$ where $\mathbf{X}_0$ is the set of Cartesian coordinates where energy is minimized. For an ensemble that is generated by small conformational changes induced by Brownian motion bombardment of a molecule that is free to move in the absence of an external potential, $V(\mathbf{X}\,;\,\mathbf{X}_0) \approx \frac{1}{2}(\mathbf{X}-\mathbf{X}_0)^T[K(\mathbf{X}_0)](\mathbf{X}-\mathbf{X}_0)$, where $[K(\mathbf{X}_0)] = [K(\mathbf{X}_0)]^T$ is the Hessian (or stiffness) matrix evaluated at $\mathbf{X}_0$, the referential (lowest energy) conformation,

$$K_{ij}(\mathbf{X}_0) = \left.\frac{\partial^2 V}{\partial X_i \partial X_j}\right|_{\mathbf{X}=\mathbf{X}_0}$$

In this case, the Maxwell-Boltzmann distribution will be

$$f(\mathbf{X},\mathbf{P}\,;\,\mathbf{X}_0,\beta) \;\doteq\; \frac{1}{Z(\beta\,;\,\mathbf{X}_0)} \exp \beta \cdot \left(-\frac{1}{2}(\mathbf{X}-\mathbf{X}_0)^T[K(\mathbf{X}_0)](\mathbf{X}-\mathbf{X}_0) - \frac{1}{2}\mathbf{P}^T[M]^{-1}\mathbf{P}\right) \tag{8}$$

where

$$Z(\beta\,;\,\mathbf{X}_0) = \int_{\mathbf{X}\in\mathbb{R}^{3n}}\int_{\mathbf{P}\in\mathbb{R}^{3n}} \exp\beta \cdot \left(-\frac{1}{2}(\mathbf{X}-\mathbf{X}_0)^T[K(\mathbf{X}_0)](\mathbf{X}-\mathbf{X}_0) - \frac{1}{2}\mathbf{P}^T[M]^{-1}\mathbf{P}\right) d\mathbf{P}\, d\mathbf{X}$$

is the partition function that normalizes $f(\mathbf{X},\mathbf{P})$ to be a probability density, and $\beta = 1/k_BT$ where k_B is the Boltzmann constant and T is temperature in degrees Kelvin. The distribution in Equation (8) is a $6n$-dimensional Gaussian distribution with covariance

$$\Sigma = \beta^{-1} \cdot ([K(\mathbf{X}_0)]^{-1} \oplus [M]) \tag{9}$$

where $\oplus$ denotes the direct sum of two matrices.

2.1. Invariance and Degeneracy

In some scenarios there is an external elastic potential corresponding to the macromolecule being tethered to a substrate, in which case Equation (9) would be valid. However, an important technical detail that needs to be addressed when there is no external potential is that $[K(\mathbf{X}_0)]$ will be singular. This is related to the fact that when there is no external potential, then for any static $g \in G$,

$$V(\mathbf{X}\,;\,\mathbf{X}_0) \;=\; V(g\cdot\mathbf{X}\,;\,g\cdot\mathbf{X}_0) \tag{10}$$

where

$$g\cdot\mathbf{X} \;\doteq\; [(g\cdot\mathbf{x}_1)^T,...,(g\cdot\mathbf{x}_n)^T]^T$$

From the above invariance of $V(\cdot\,;\,\cdot)$ under rigid-body shifts it follows that

$$[K((R\oplus R\oplus\cdots\oplus R)\mathbf{X}_0)] = (R\oplus R\oplus\cdots\oplus R)\,[K(\mathbf{X}_0)]\,(R\oplus R\oplus\cdots\oplus R)^T \tag{11}$$

Moreover,

$$V(g^{-1}\cdot\mathbf{X}\,;\,\mathbf{X}_0) \;=\; V(\mathbf{X}\,;\,g\cdot\mathbf{X}_0) \tag{12}$$

This means that

$$f(g\cdot\mathbf{X},\mathbf{P}\,;\,g\cdot\mathbf{X}_0) = f(\mathbf{X},\mathbf{P}\,;\,\mathbf{X}_0)$$

It is also true that

$$f(\mathbf{X},R\cdot\mathbf{P}\,;\,\mathbf{X}_0) = f(\mathbf{X},\mathbf{P}\,;\,\mathbf{X}_0)$$

where $R\cdot\mathbf{P} = [(R\mathbf{p}_1)^T,...,(R\mathbf{p}_n)^T]^T$, but this fact is not relevant to the current discussion.

In classical statistical mechanics, this invariance under translation and rotation is broken by limiting the motion of molecules to a finite box, which is tantamount to defining an external potential. But in the present context it will be convenient instead to reduce degrees of freedom while maintaining the Gaussian nature of this distribution (which is convenient for the calculations that will follow). The mass matrix has full rank, and so R is not a concern. But the stiffness matrix has rank $3n-6$ due to the 6 degrees of freedom in G. Therefore, either reducing $[K]$ by six degrees of freedom, or by adding an artificial tethering potential, will make it have full rank. If $[K]$ is decomposed as

$$[K] = Q(\mathbb{O}_6\oplus\Lambda)Q^T$$

where $\mathbb{O}_6$ is the 6×6 matrix of zeros and Λ is the $3n-6$-dimensional diagonal matrix with the remaining eigenvalues, then choosing

$$\mathbf{X} = Q\begin{pmatrix}\mathbf{0}_6\\ \mathbf{Y}\end{pmatrix}$$

will define a new (full-rank) stiffness matrix, which is Λ, where $\mathbf{0}_6$ is the six-dimensional zero vector.

Alternatively, constraints

$$\sum_{i=1}^{n} m_i(\mathbf{x}_i-\mathbf{x}_i^0) = \mathbf{0} \quad\text{and}\quad \sum_{i=1}^{n}\mathbf{x}_i^0\times(m_i\mathbf{x}_i) = \mathbf{0}$$

can be imposed to ensure that the body does not undergo any overall rotation or translation from its initial state as it deforms. These result from conservation of linear and angular momentum. They can be imposed as hard constraints, by for example, solving for $\mathbf{x}_{n-1}$ and $\mathbf{x}_n$ in terms of all other $\mathbf{x}_i$'s and back substituting to produce $\tilde{\mathbf{X}}$ that is $(3n-6)$-dimensional, or as soft constraints imposed with elastic potentials. We take the approach of hard constraints. Another way to impose hard constraints would be to fix a specific point, for example, by translating every point by $-\mathbf{x}_n$ so that $\mathbf{x}_n$ becomes $\mathbf{0}$, and keep it fixed there as the macromolecule undergoes fluctuations. This removes three degrees of freedom. Then, if the vector from $\mathbf{x}_n$ to $\mathbf{x}_{n-1}$ is forced to have a specific orientation (e.g., pointing along the direction $\mathbf{e}_1 = [1,0,0]^T$), this removes two more degrees of freedom. The final rigid-body degree of freedom can be enforced by defining allowable motions of $\mathbf{x}_{n-2}$ to be in the plane spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$. Additional information on applying these constraints can be found in Appendix A.

2.2. Conformational Boltzmann Distribution

We shall be concerned primarily with the conformational Boltzmann distribution, which results from removing the superfluous rigid-body degrees of freedom that leave the potential invariant as in Equations (10) and (12). Removing these degrees of freedom and marginalizing over the conjugate momenta results in

$$f_c(\tilde{\mathbf{X}}\,;\,\tilde{\mathbf{X}}_0,\beta) \doteq \frac{1}{Z_c(\beta\,;\,\tilde{\mathbf{X}}_0)} \exp\left(-\frac{1}{2}\beta(\tilde{\mathbf{X}}-\tilde{\mathbf{X}}_0)^T[\tilde{K}(\tilde{\mathbf{X}}_0)](\tilde{\mathbf{X}}-\tilde{\mathbf{X}}_0)\right) \tag{13}$$

where

$$Z_c(\beta\,;\,\tilde{\mathbf{X}}_0) = \int_{\tilde{\mathbf{X}}\in\mathbb{R}^{3n-6}} \exp\left(-\frac{1}{2}\beta(\tilde{\mathbf{X}}-\tilde{\mathbf{X}}_0)^T[\tilde{K}(\tilde{\mathbf{X}}_0)](\tilde{\mathbf{X}}-\tilde{\mathbf{X}}_0)\right)\, d\tilde{\mathbf{X}}$$

Here $\tilde{X}$ represents a reduced set of coordinates that could have been generated in either of the two ways discussed above. Now that $[\tilde{K}(\tilde{\mathbf{X}}_0)]$ is full rank, the conformational partition function can be computed in closed form as

$$Z_c(\beta\,;\,\tilde{\mathbf{X}}_0) = \left(\frac{2\pi}{\beta}\right)^{\frac{3n-6}{2}} \left(\det[\tilde{K}(\tilde{\mathbf{X}}_0)]\right)^{-\frac{1}{2}}$$

2.3. Elastic Network Models

Elastic network models have been used to model the motion of macromolecules [28–35]. In the elastic network model the stiffness matrix is computed from a set of coordinates $\{\mathbf{x}_i^0\}$ of a baseline structure in terms of 3×3 blocks as

$$K_{ij} = \begin{cases} -G_{ij} & \text{for} \quad i \neq j \\ \sum_{1\leq k\neq i\leq n} G_{k,i} & \text{for} \quad i = j \end{cases} \tag{14}$$

where

$$G_{ij} = k_{ij}\frac{(\mathbf{x}_i^0-\mathbf{x}_j^0)(\mathbf{x}_i^0-\mathbf{x}_j^0)^T}{\|\mathbf{x}_i^0-\mathbf{x}_j^0\|^2}$$

and k_{ij} takes a uniform nonzero value when $\|\mathbf{x}_i^0-\mathbf{x}_j^0\|$ is within a cutoff radius, r_0, and takes a value of zero otherwise. This leads to a stiffness matrix that is sparse, which leads to efficient computations of normal modes. For our purposes, it will be more convenient to define k_{ij} as follows:

$$k_{ij} = k\cdot\phi(\|\mathbf{x}_i^0-\mathbf{x}_j^0\|)$$

where $\phi(0) = 1$ and $\phi(r) \approx 1$ when $0 < r < r_0$, and decays rapidly to zero on the range r_0 to infinity. Many such functions exist that are differentiable, including sigmoid functions and certain multi-Gaussian difference functions. Then the full stiffness matrix constructed from these blocks explicitly depends on the vector of initial coordinates as $[K] = [K(\mathbf{X}_0)]$.

3. Similarity Measures for Conformational Ensembles

In this section, measures of similarity that build on the concepts in the previous section are articulated. Let the two ensembles in question be described by probability density functions $f_1(\tilde{\mathbf{X}}) = f_c(\tilde{\mathbf{X}}\,;\,\tilde{\mathbf{X}}_0, \beta)$ and $f_2(\tilde{\mathbf{X}}) = f_c(\tilde{\mathbf{X}}\,;\,\tilde{\mathbf{X}}_0', \beta')$. The difference between them is that either $\tilde{\mathbf{X}}_0 \neq \tilde{\mathbf{X}}_0'$, or $\beta \neq \beta'$, or both. Since the conformations were effectively anchored in space in a specific way prior to computing these probability densities, in order to make the stiffness matrix nonsingular, it follows that the probability densities should be optimally aligned before they are compared. This is analogous to what was done in Equation (5) in comparing shapes using their mass densities. But now, the conformational probability densities are functions on $(3n-6)$-dimensional Euclidean space. Herein several measures are compared. Throughout this discussion,

$$N \doteq 3n - 6$$

Let $\mathcal{P}(\mathbb{R}^N)$ denote the set of all smooth probability density functions on $\mathbb{R}^N$. Given a measure of distance, divergence, or discrepancy between two probability density functions,

$$D : \mathcal{P}(\mathbb{R}^N) \times \mathcal{P}(\mathbb{R}^N) \longrightarrow \mathbb{R}_{\geq 0}$$

and given an action of G on $\mathbb{R}^N$, $g \cdot \mathbb{R}^N$, a corresponding action on functions can be defined in analogy with Equation (6). A measure of similarity between probability densities describing an ensemble can then be constructed as

$$\Delta(f_1\,;\,f_2) \doteq \min_{g \in G} D(f_1\,;\,g \cdot f_2) \tag{15}$$

This similarity of conformational probability densities then serves to quantify how similar two ensembles are. Below, it is shown how various $D(f_1\,;\,f_2)$ can be computed in closed form when f_1 and f_2 are Gaussian distributions.

3.1. L^2 Difference

It is known from the literature that the following measure can be computed in closed form for Gaussian distributions on $\mathbb{R}^N$, $f_{(\boldsymbol{\mu},\Sigma)}(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$:

$$\begin{aligned} D_2\left(f_{(\boldsymbol{\mu}_1,\Sigma_1)}\,,\,f_{(\boldsymbol{\mu}_2,\Sigma_2)}\right) &= \left(\int_{\mathbb{R}^N} [f(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1) - f(\mathbf{x}; \boldsymbol{\mu}_2, \Sigma_2)]^2 d\mathbf{x}\right)^{\frac{1}{2}} \qquad (16) \\ &= \left(\int_{\mathbb{R}^N} f(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1)^2 - 2\, f(\mathbf{x}; \boldsymbol{\mu}_1, \Sigma_1) f(\mathbf{x}; \boldsymbol{\mu}_2, \Sigma_2) + f(\mathbf{x}; \boldsymbol{\mu}_2, \Sigma_2)^2 d\mathbf{x}\right)^{\frac{1}{2}} \end{aligned}$$

An advantage of Gaussian functions is that the integration of quadratic terms over $\mathbb{R}^N$ has a closed-form expression. We derived it as follows,

$$\begin{aligned}\int_{\mathbb{R}^N} f(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)f(\mathbf{x};\boldsymbol{\mu}_2;\Sigma_2)d\mathbf{x} &= \int_{\mathbb{R}^N}(2\pi)^{-N/2}|\det\Sigma_1|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right)\\ &\quad (2\pi)^{-N/2}|\det\Sigma_2|^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right)d\mathbf{x}\\ &=(2\pi)^{-N}|\det\Sigma_1\det\Sigma_2|^{-1/2}\int_{\mathbb{R}^N}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right.\\ &\quad\left.-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right)d\mathbf{x}\end{aligned}\tag{17}$$

Since

$$\int_{\mathbb{R}^N}\exp(-\frac{1}{2}\mathbf{x}^TM\mathbf{x}-m^T\mathbf{x}-C)d\mathbf{x}=(2\pi)^{N/2}|\det M|^{-1/2}\exp(\frac{1}{2}m^TM^{-1}m-C)\tag{18}$$

Equation (17) can be rewritten in a closed-form as

$$\begin{aligned}\int_{\mathbb{R}^N} f(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)f(\mathbf{x};\boldsymbol{\mu}_2;\Sigma_2)d\mathbf{x} =&(2\pi)^{-N/2}\left|\det\Sigma_1\det\Sigma_2\det(\Sigma_1^{-1}+\Sigma_2^{-1})\right|^{-1/2}\\ &\exp\left(\frac{1}{2}(\boldsymbol{\mu}_1^T\Sigma_1^{-1}+\boldsymbol{\mu}_2^T\Sigma_2^{-1})(\Sigma_1^{-1}+\Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\boldsymbol{\mu}_1+\Sigma_2^{-1}\boldsymbol{\mu}_2)\right.\\ &\left.-\frac{1}{2}(\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\mu}_2^T\Sigma_2^{-1}\boldsymbol{\mu}_2)\right)\end{aligned}\tag{19}$$

Similarly, we can write

$$\int_{\mathbb{R}^N}\left(f(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)\right)^2d\mathbf{x}=2^{-N}\pi^{-N/2}\left|\det\Sigma_1\right|^{-1/2}\tag{20}$$

For the numerical examples presented in Section 4 a normalized version of the L^2 difference is presented. This normalized version is given by

$$\overline{D}_2\left(f_{(\boldsymbol{\mu}_1,\Sigma_1)}\,,\,f_{(\boldsymbol{\mu}_2,\Sigma_2)}\right)=\frac{D_2\left(f_{(\boldsymbol{\mu}_1,\Sigma_1)}\,,\,f_{(\boldsymbol{\mu}_2,\Sigma_2)}\right)}{\left(\int_{\mathbb{R}^N}[f(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)]^2d\mathbf{x}\right)^{\frac{1}{2}}}\tag{21}$$

3.2. Kullback-Leibler Divergence

The Kullback-Leibler divergence (or relative entropy) between two probability density functions is defined as

$$D_{KL}\left(f_{(\boldsymbol{\mu}_1,\Sigma_1)}\,\|\,f_{(\boldsymbol{\mu}_2,\Sigma_2)}\right)\;=\;\int_{\mathbb{R}^N}f(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)\log\frac{f(\mathbf{x};\boldsymbol{\mu}_1,\Sigma_1)}{f(\mathbf{x};\boldsymbol{\mu}_2,\Sigma_2)}d\mathbf{x}\tag{22}$$

This can be computed in closed form for Gaussians using the identity in Equation (18) and

$$\int_{\mathbb{R}^N}\mathbf{x}^TG\mathbf{x}\exp\left(-\frac{1}{2}\mathbf{x}^TA\mathbf{x}\right)d\mathbf{x}=(2\pi)^{N/2}\frac{\mathrm{tr}(GA^{-1})}{|\det A|^{\frac{1}{2}}}\tag{23}$$

as follows:

$$D_{KL}(f_{(\boldsymbol{\mu}_1,\Sigma_1)}||f_{(\boldsymbol{\mu}_2,\Sigma_2)})=\frac{1}{2}\log\frac{|\det\Sigma_2|}{|\det\Sigma_1|}-\frac{N}{2}+\frac{1}{2}\left[\mathrm{tr}(\Sigma_2^{-1}\Sigma_1)+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right]\tag{24}$$

4. Numerical Examples and Comparison with RMSD

The examples presented below use structural data from the protein database (PDB) [36] and the 2010 Critical Assessment of protein Structure Prediction (CASP9) [37] to highlight how the metrics in Section 3 differ with respect to RMSD.

4.1. Example 1: Perturbed Protein Structure

In order to contrast the proposed metrics in Section 3 with those that are purely geometric based, we perturbed a conformation given in the PDB [36] in two ways. The protein chosen is Streptomyces griseus protease B (PDB ID 3SGB) [36,38] shown in Figure 1. This protein's loops and size make it relatively flexible and well suited to illustrate the difference between these metrics.

Figure 1. Biological assembly image of the example protein used, Streptomyces griseus protease B. Image from the RCSB PDB [39] of PDB ID 3SGB [36,38].

The perturbations are performed using a unit length vector, $\mathbf{V} \in \mathbb{R}^N$. Using $\mathbf{V}$ and the alpha carbon locations given in the PDB file, $\mathbf{X}_0^0$, a family of similar mean conformations were generated as

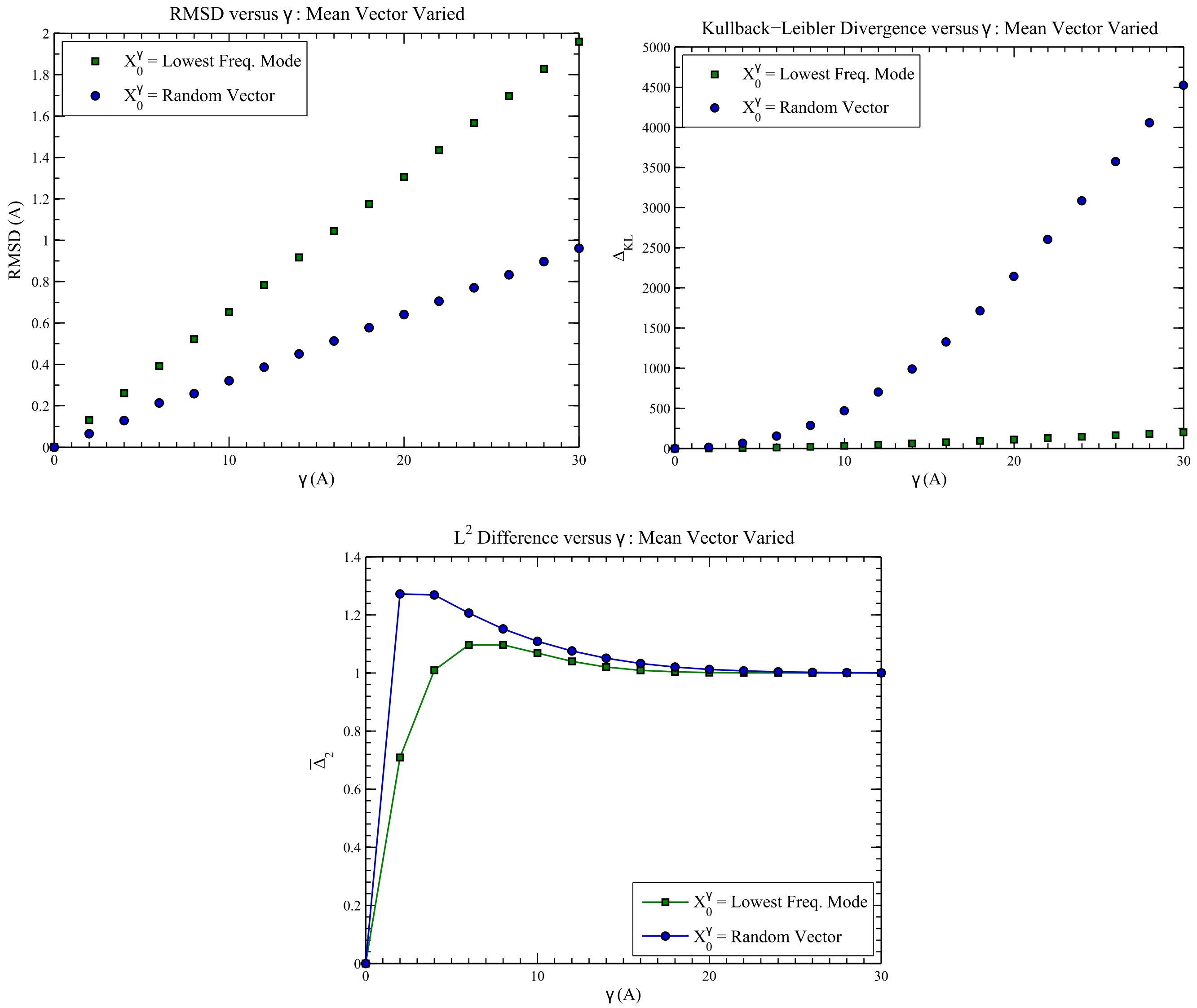
$$\mathbf{X}_0^\gamma = \mathbf{X}_0^0 + \gamma \mathbf{V} \tag{25}$$

for scalar values of γ. This parameterization of mean conformations then provides a way to compare how the difference measures given in Section 3 vary over a range of RMSD values. Conformational ensembles were then generated using the conformational Boltzmann distribution described in Section 2.2 together with the hard constraints discussed in Sections 2.1 and Appendix A.

The first set of comparisons that was performed was between two different perturbation vectors, $\mathbf{V}$. For the first vector we consider perturbations in the direction of low frequency normal modes. These mode shapes are associated with the dominant direction of motion for large conformational changes. We start by forming a coarse grain elastic network as described in Section 2.3 using the alpha carbons. Based on this elastic network, normal mode analysis was used to obtain the minimum nonzero eigenvalue and its associated unit-length eigenvector. This eigenvector is then used as the perturbation vector $\mathbf{V}$. The second type of perturbation applied was obtained by taking a random vector whose elements are drawn from a zero mean normal distribution with constant variance. This vector was then normalized to unit length. For these examples, all stiffness matrices were constructed using $r_0 = 8$ Å, $k = 1$, and $\beta = 1$.

Figure 2 highlights the difference of these two perturbation methods. The conformations corresponding to the random perturbation exhibit lower RMSD values at all values of γ than perturbation in the likely direction of motion; for the random perturbations, the RMSD values are roughly half of those associated with the mode shape. This is particularly interesting when contrasted with the Kullback-Leibler divergences and L^2 differences. For these measures, the values associated with the normal mode perturbations are lower than those associated with the random perturbations. This can be attributed to the fact these deviations occur in a direction of high flexibility and both of these measures are functions of the conformation's elastic structure. This effect is particularly pronounced for the Kullback-Leibler divergence. This result highlights the fact that RMSD alone does not capture how similar two conformations are when uncertainty is considered.

Figure 2. Here we demonstrate how RMSD (top left), Kullback-Leibler divergence (top right), and normalized L^2 difference (bottom) vary with γ (β and r_0 are fixed at 1 and 8 respectively). Two perturbation vectors are given, one corresponding to the lowest frequency mode and one random unit vector.



The second set of comparisons presented utilizes only the normal mode perturbation. However, a variety of values for β and r_0 were used as these parameters effectively alter the "stiffness" of the conformational Boltzmann distribution for a fixed value of k ($k = 1$ for these examples). Figures 3 and 4 demonstrate the effect of varying β and the cutoff radius, r_0. It should be noted that the wide range of β values being shown are only for the purpose of illustrating the relative sensitivity of our results to β. The values used may not represent a physiologically realizable range.

Figure 3. These plots illustrate how Kullback-Leibler divergence, $\Delta_{KL}(f_c(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_0^0, \beta) \| f_c(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_0^\gamma, \beta))$ varies with γ for several values of β (left) and several values of r_0 (right).

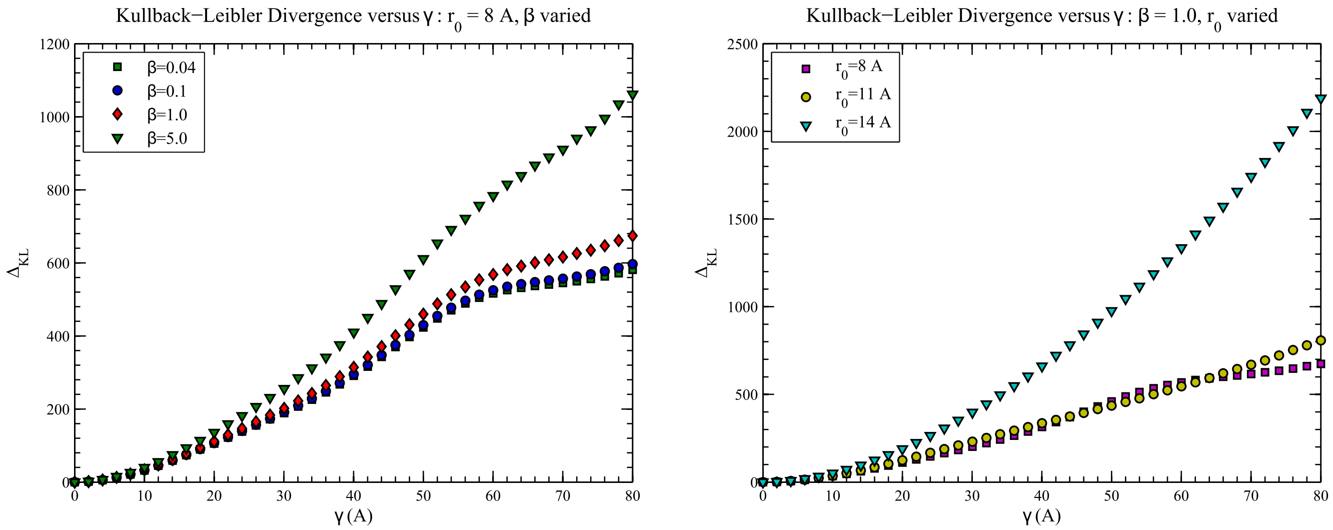
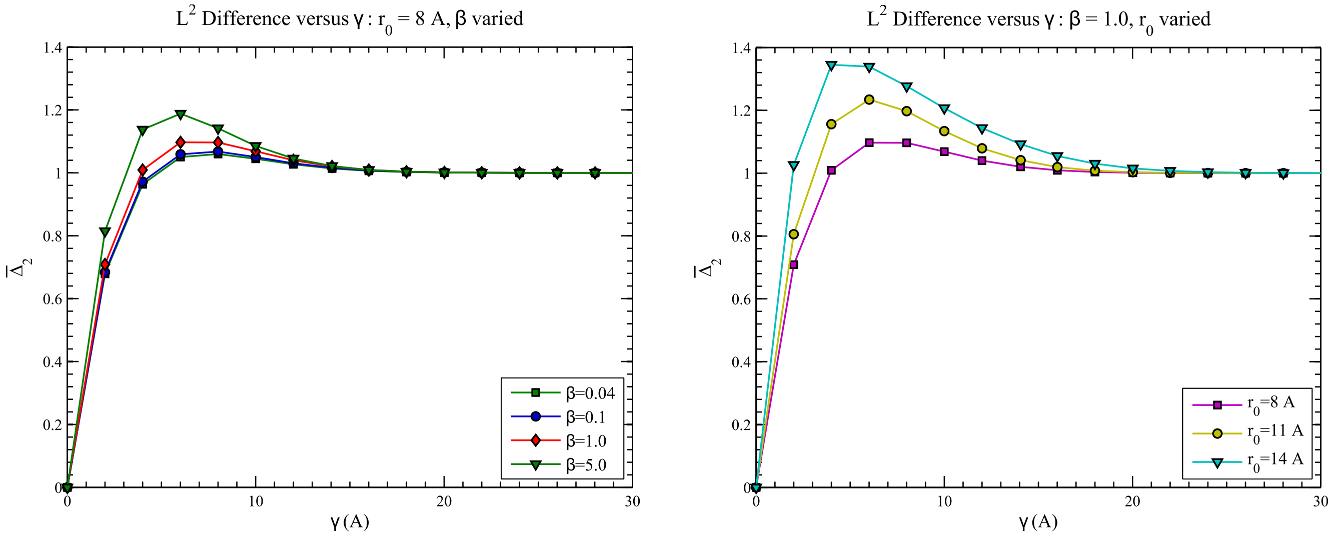


Figure 4. These plots illustrate how normalized L^2 difference, $\overline{\Delta}_2(f_c(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_0^0, \beta), f_c(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}_0^\gamma, \beta))$ varies with γ for several values of β (left) and several values of r_0 (right).



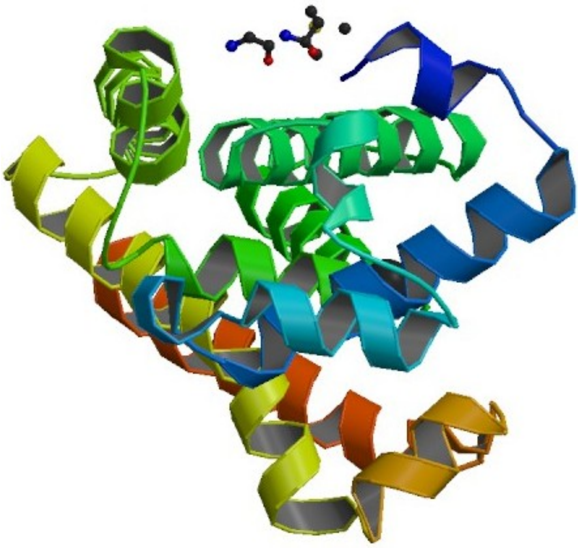
These figures illustrate several interesting things about how these metrics differ from RMSD. The first thing to note is that while RMSD is constant with respect to β and r_0, both the Kullback-Leibler divergence and L^2 difference vary with respect to these flexibility measures. It is also apparent that Δ_{KL} increases with γ; although, the relationship is not linear. For "stiffer" structures (*i.e.*, higher β or higher r_0), the slope of Δ_{KL} with respect to γ is lower and begins to flatten sooner. The relationship between

$\overline{\Delta}_2$ and γ is more complex. Close to $\gamma = 0$, $\overline{\Delta}_2$ increases with γ until it reaches a maximum and then begins to decline until it levels out at a value of one.

4.2. Example 2: CASP Predictions

One of the motivations for looking at metrics for conformational ensembles is for comparison of protein predictions to experimentally solved proteins. Here we present a target protein structure, T0575 from CASP9 [37] (PDB ID 3NRG [40]), and several groups' predictions. Figure 5 provides a ribbon diagram for the target protein. We examined the 44 predictions of the target that met the following conditions: (1) the prediction contains all of the residues; (2) the prediction is a group's "first" model (*i.e.*, it is the model with which a group is most confident); and (3) the RMSD between the alpha carbons of the prediction and target is less than 6 Å. For each of the predictions and the target, ensembles were generated using the conformational Boltzmann distribution for $\beta = 1$ and $r_0 = 8$ Å. Using these ensembles, the Kullback-Leibler divergence Δ_{KL} was determined. Figure 6 presents these values of Kullback-Leibler divergence versus their associated RMSD.

Figure 5. Biological assembly image of the protein target used, a TetR family transcriptional regulator (Caur_2714) from Chloroflexus aurantiacuss. Image from the RCSB PDB [39] of PDB ID 3NRG [36,40].



It is believed that the Kullback-Leibler divergence in this context provides an indication of how similar two conformations are in terms of topology and flexibility. To demonstrate this, we used normal mode analysis to perturb the predicted models; we added a weighted sum of the first few nonzero normal modes to the predicted models to minimize the RMSD between the prediction and the target. These first few nonzero normal modes represent the primary directions of motion for elastic networks. Let the vectors associated with the normal modes be ordered such that $\mathbf{V}_i$ represents the i^{th} nonzero normal mode. Then we can represent this minimization over m normal modes with the following perturbed model vector similar to what was done in Equation (25)

$$\mathbf{X}_0^{\gamma} = \mathbf{X}_0^0 + \sum_{i=1}^{m} \gamma_i \mathbf{V}_i \tag{26}$$

where $\mathbf{X}_0^0$ is a vector of the original coordinates of the predicted model and γ_i's are scalar weights. If $\mathbf{X}_0^{Target}$ is the coordinate vector associated with the target structure, we can then use a cyclic gradient descent to minimize RMSD,

$$\min_{\gamma_i} d_2(\mathbf{X}_0^{Target}, \mathbf{X}_0^{\gamma}) \tag{27}$$

The results for this analysis can found in Figures 6 and 7 for the first three normal modes ($m = 3$). Figure 6 shows the original Kullback-Leibler divergence versus both the original RMSD and the RMSD optimized using the first three normal modes. Figure 7 provides the percent reduction in RMSD versus the original Kullback-Leibler divergence. In this analysis, we would expect to see the amount of minimization decrease as the Kullback-Leibler divergence increases. This trend is apparent, although there are a few outliers. We note that similar results were observed for other CASP9 targets and different values of m.

Figure 6. The Kullback-Leibler divergences between predicted models and the CASP9 target are shown with respect to RMSD. RMSDs are given for the original predictions and predictions perturbed along the first three normal modes.

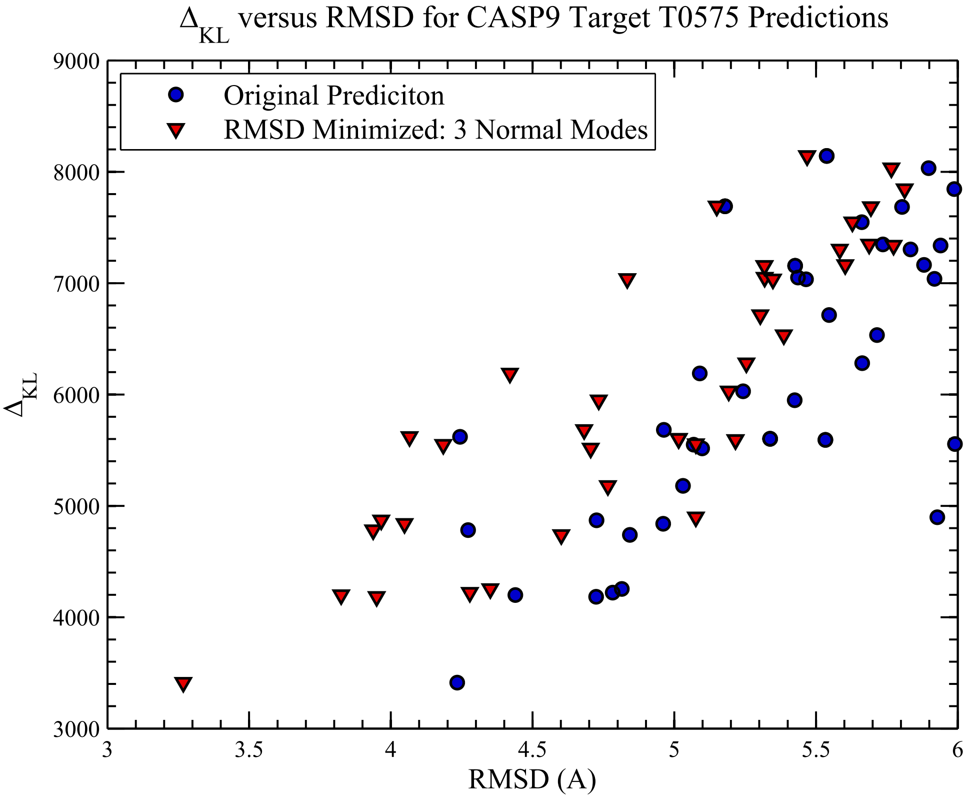
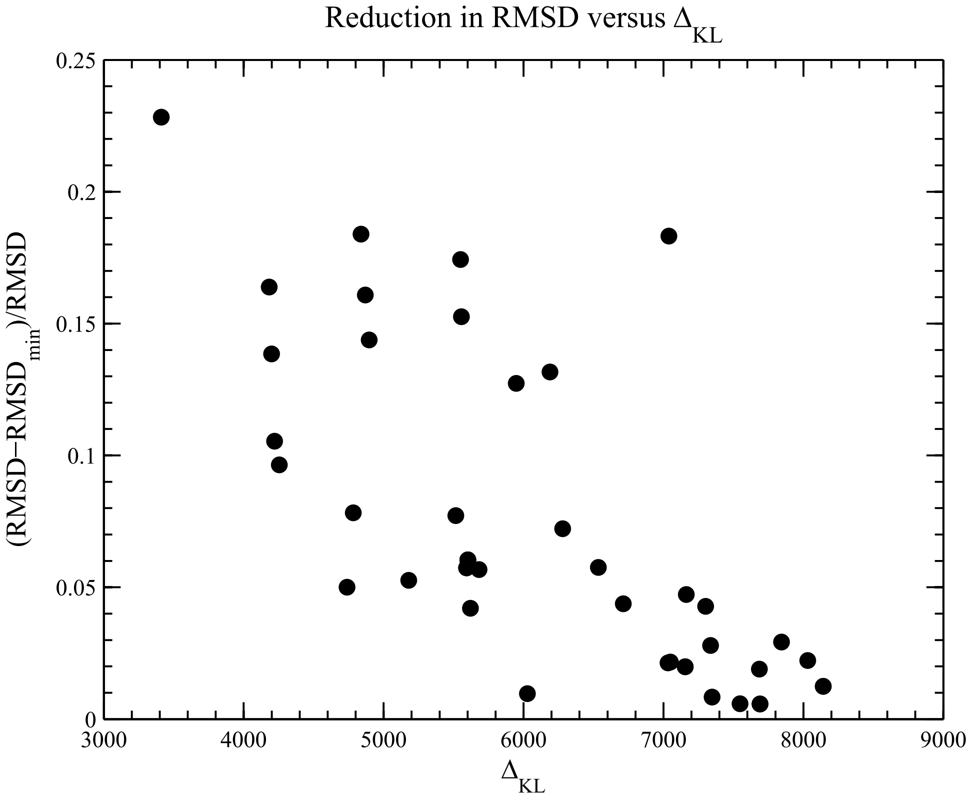


Figure 7. The reduction in RMSD realized by minimizing along the first three normal modes is given with respect to the Kullback-Leibler divergence between the predicted model and the target (T0575).



5. Conclusions

Measures to quantify how similar two conformational ensembles are have been articulated here. In contrast to well-known measures of shape similarity between two conformations, which are purely geometric, the comparison of ensembles here requires concepts from probability and information theory. We have illustrated how using the stiffness information in conformational ensembles for two macromolecules can be used to develop difference measures that take into account the positional uncertainty of flexible regions. These new measures are based on the Kullback-Leibler divergence and L^2 difference.

Numerical examples of how these measures compare with the traditional RMSD were presented. The conformational ensembles for these examples were based on coarse-grained elastic networks. It has been demonstrated that for two different conformations, the RMSD of each of them with respect to a third may not provide enough information to determine which of the conformations is more similar to the third in terms of their positional uncertainty.

The measures presented provide additional information with respect to positional uncertainty. They have the potential to be extended or combined with each other, additional information theoretic measures, and/or traditional geometric measures to make them more robust to a variety of parameters. Additionally, methods for comparing the similarity of an individual structure and an ensemble can be explored.

Acknowledgements

The authors would like to thank Ms. Yan Yan and Mr. M. Kendall Ackerman for their insights and helpful comments. This research was partially supported under an appointment to the Department of Homeland Security (DHS) Scholarship and Fellowship Program, administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and DHS. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05-06OR23100. This work was also supported in part by NSF grant IIS-0915542 RI: Small: Robotic Inspection, Diagnosis, and Repair.

References and Notes

1. Holm, L.; Sander, C. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **1993**, *233*, 123–138.
2. Mitchell, E.M.; Artymiuk, P.J.; Rice, D.W.; Willett, P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **1989**, *212*, 151–166.
3. Zemla, A.; Venclovas, C.; Moult, J.; Fidelis, K. Processing and analysis of CASP3 protein structure predictions. *Protein. Struct. Funct. Bioinform.* **1999**, *37*, 22–29.
4. Gong, H.P.; Rose, G.D. Does secondary structure determine tertiary structure in proteins? *Protein. Struct. Funct. Bioinform.* **2005**, *61*, 338–343.
5. Irving, J.A.; Whisstock, J.C.; Lesk, A.M. Protein structural alignments and functional genomics. *Proteins* **2001**, *42*, 378–382.
6. Jewett, A.I.; Huang, C.C.; Ferrin, T.E. MINRMS: An efficient algorithm for determining protein structure similarity using root-mean-squared-distance. *Bioinformatics* **2003**, *19*, 625–634.
7. Hilser, V.J.; Garcia-Moreno, E.B.; Oas, T.G; Kapp, G.; Whitten, S.T. A statistical thermodynamic model of the protein ensemble. *Chem. Rev.* **2006**, *106*, 1545–1558.
8. Hilser, V.J.; Dowdy, D.; Oas, T.G.; Freire, E. The structural distribution of cooperative interactions in proteins: Analysis of the native state ensemble. *Proc. Nat. Acad. Sci. U. S. A.* **1998**, *95*, 9903–9908.
9. Tyka, M.D.; Keedy, D.A.; Andre, I.; Dimaio, F.; Song, Y.; Richardson, D.C.; Richardson, J.S.; Baker, D. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Struct. Biol.* **2011**, *405*, 607–618.
10. Burra, P.V.; Zhang, Y.; Godzik, A.; Stec, B. Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. *Proc. Nat. Acad. Sci. U. S. A.* **2009**, *106*, 10505–10510.
11. Chirikjian, G.S. Modeling loop entropy. *Methods Enzymol.* **2011**, *487*, 99–132.
12. Zhou, H.-X. Loops in proteins can be modeled as worm-like chains. *J. Phys. Chem. B* **2001**, *105*, 6763–6766.
13. Zhang, J.; Lin, M.; Chen, R.; Wang, W.; Liang, J. Discrete state model and accurate estimation of loop entropy of RNA secondary structures. *J. Chem. Phys.* **2008**, *128*, doi:10.1063/1.2895050.
14. Shehu, A.; Clementi, C.; Kavraki, L.E. Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. *Protein. Struct. Funct. Bioinform.* **2006**, *65*, 164–179.

15. Shehu, A.; Kavraki, L.E.; Clementi, C. On the characterization of protein native state ensembles. *Biophys. J.* **2007**, *92*, 1503–1511.
16. Wu, X.; Brooks, B.R. Toward canonical ensemble distribution from self-guided Langevin dynamics simulation. *J. Chem. Phys.* **2011**, *134*, 134108:1–134108:12.
17. Grosberg, A.Y.; Khokhlov, A.R. *Statistical Physics of Macromolecules*; American Institute of Physics: New York, NY, USA, 1994.
18. Jacobs, D.J.; Rader, A.J.; Kuhn, L.A.; Thorpe, M.F. Protein flexibility predictions using graph theory. *Protein. Struct. Funct. Bioinform.* **2001**, *44*, 150–165.
19. Damm, K.L.; Carlson, H.A. Gaussian-weighted RMSD superposition of proteins: A structural comparison for flexible proteins and predicted protein structures. *Biophys. J.* **2006**, *90*, 4558–4573.
20. Moult, J.; Pederson, J.T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Protein. Struct. Funct. Genet.* **1995**, *23*, ii–iv.
21. Moult, J.; Fidelis, K.; Kryshtafovych, A.; Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—Round IX. *Protein. Struct. Funct. Bioinform.* **2011**, *79*, 1–5.
22. Wodak, S.J. From the Mediterranean coast to the shores of Lake Ontario: CAPRI's premiere on the American continent. *Protein. Struct. Funct. Bioinform.* **2007**, *69*, 697–698.
23. Janin, J. Welcome to CAPRI: A critical assessment of predicted interactions. *Protein. Struct. Funct. Bioinform.* **2002**, *47*, doi:10.1002/prot.10111.
24. Here the word "similarity" is not used in the sense of basic geometry which involves rigid-body and scale transformations, but rather describes small differences in shape.
25. Chiang, C.-J.; Chirikjian, G.S. Similarity metrics with applications in modular robot motion planning. *Auton. Robot.* **2001**, *10*, 91–106.
26. Pamecha, A.; Ebert-Uphoff, I.; Chirikjian, G.S. Useful metrics for modular robot motion planning. *IEEE Trans. Robot. Autom.* **1997**, *13*, 531–545.
27. The interpretation of the "action" $\cdot$ depends on the context. We can talk about the action on a function, $g \cdot \rho'$, or on a vector, $g \cdot \mathbf{x}$, or other objects.
28. 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*, 173–180.
29. Kim, M.K.; Chirikjian, G.S.; Jernigan, R.L. Elastic models of conformational transitions in macromolecules. *J. Mol. Graph. Model.* **2002**, *21*, 151–160.
30. Zheng, W.; Brooks, B.R.; Hummer, G. Protein conformational transitions explored by mixed elastic network models. *Protein. Struct. Funct. Bioinform.* **2007**, *69*, 43–57.
31. Maragakis, P.; Karplus, M. Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. *J. Mol. Biol.* **2005**, *352*, 807–822.
32. 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*, 505–515.
33. Bahar, I.; Atilgan, A.R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Design* **1997**, *2*, 173–181.
34. Doruker, P.; Jernigan, R.L.; Bahar, I. Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.* **2002**, *23*, 119–127.

35. Wang, Y.; Rader, A.J.; Bahar, I.; Jernigan, R.L. Global ribosome motions revealed with elastic network model. *J. Struct. Biol.* **2004**, *147*, 302–314.
36. 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*, 235–242.
37. CASP9. Protein Structure Prediction Center, University of California, Davis.2010. Available online: http://predictioncenter.org/casp9/ (accessed on 25 January 2012).
38. Read, R.J.; Fujinaga, M.; Sielecki, A.R.; James, M.N. Structure of the complex of Streptomyces griseus protease B and the third domain of the turkey ovomucoid inhibitor at 1.8-A resolution. *Biochemistry* **1983**, *22*, 4420–4433, PDB ID: 3GSB.
39. RCSB Protein Data Bank. Available online: http://www.pdb.org (accessed on 20 January 2012).
40. Joint Center for Structural Genomics. Crystal structure of a TetR family transcriptional regulator (Caur_2714) from Chloroflexus aurantiacus J-10-FL at 2.56 A resolution. PDB ID: 3NRG.

Appendix

A. Restricting the Conformational Boltzmann Distribution and Applying a Rigid-Body Transformation

Recall from Section 2.1 that we can remove 6 degrees of freedom by forcing $\mathbf{x}_n^0 = \mathbf{0}$, rotating the conformation so that the vector from $\mathbf{x}_n^0$ to $\mathbf{x}_{n-1}^0$ is aligned with $\mathbf{e}_1$, and confining $\mathbf{x}_{n-2}^0$ to the plane spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$. The resulting coordinate vector $\mathbf{X} \in \mathbb{R}^{3n}$ can then be written as

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{n-3} \\ \mathbf{x}_{n-2} \\ \mathbf{x}_{n-1} \\ \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{n-3} \\ a_1\mathbf{e}_1 + a_2\mathbf{e}_2 \\ a_3\mathbf{e}_1 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{n-3} \\ a_1 \\ a_2 \\ 0 \\ a_3 \\ 0 \\ 0 \\ \mathbf{0} \end{pmatrix} \tag{28}$$

Let us define $\mathbf{B}$ as

$$\mathbf{B} = \begin{pmatrix} \mathbb{I}_{(3n-7)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{29}$$

Here $\mathbf{B} \in \mathbb{R}^{(3n-6)\times 3n}$. Let us also define

$$\mathbf{X}_{1,n-3} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{n-3} \end{pmatrix} \tag{30}$$

The generalized coordinate system $\tilde{\mathbf{X}} \in \mathbb{R}^{3n-6}$ is then

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{n-3} \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{1,n-3} \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \mathbf{B}\mathbf{X} \tag{31}$$

Here, and henceforth, a_1 will represent the signed distance of $\mathbf{x}_{n-1}$ from the plane spanned by the global $\mathbf{e}_2$ and $\mathbf{e}_3$ axes as measured along the vector from $\mathbf{x}_n^0$ to $\mathbf{x}_{n-1}^0$. This ensures that the confined degrees of freedom are respected. Also, a_2 (a_3) will represent the signed distance of $\mathbf{x}_{n-2}$ from the plane spanned by the fixed global $\mathbf{e}_2$ and $\mathbf{e}_3$ ($\mathbf{e}_1$ and $\mathbf{e}_3$) axes as measured along the transformed $\mathbf{e}_1$ ($\mathbf{e}_2$) axis.

If we apply an arbitrary rigid body motion g_0 to the conformation where

$$g_0 = \begin{pmatrix} R_0 & \mathbf{t}_0 \\ \mathbf{0}^T & 1 \end{pmatrix} \tag{32}$$

then

$$\mathbf{y}_i = R_0\mathbf{x}_i + \mathbf{t}_0 \tag{33}$$

We can then write

$$\tilde{\mathbf{R}}_0 = R_0 \oplus R_0 \oplus \cdots \oplus R_0 \tag{34}$$

and

$$\tilde{\mathbf{t}}_0 = \begin{pmatrix} \mathbf{t}_0 \\ \mathbf{t}_0 \\ \vdots \\ \mathbf{t}_0 \end{pmatrix} \tag{35}$$

where $\tilde{\mathbf{R}}_0 \in \mathbb{R}^{3n\times 3n}$ and $\tilde{\mathbf{t}}_0 \in \mathbb{R}^{3n}$. (Note: We will let $\tilde{\mathbf{R}}_0^{n-3} \in \mathbb{R}^{3(n-3)\times 3(n-3)}$ represent a direct sum of $n-3$ copies of R_0.) Then the new coordinate vector can be written as

$$\mathbf{Y} = \widetilde{\mathbf{R}}_0\,\mathbf{X} + \widetilde{\mathbf{t}}_0 \tag{36}$$

If $\mathbf{K}_{X_0}$ is the stiffness matrix associated with $\mathbf{X}$, the new stiffness matrix $\mathbf{K}_{Y_0}$ can be written as

$$\mathbf{K}_{Y_0} = \widetilde{\mathbf{R}}_0\,\mathbf{K}_{X_0}\,\widetilde{\mathbf{R}}_0^T \tag{37}$$

Also, the stiffness matrix corresponding to $\tilde{\mathbf{X}}$ can then be written as

$$\tilde{\mathbf{K}}_{X_0} = \mathbf{B}\,\mathbf{K}_{X_0}\,\mathbf{B}^T \tag{38}$$

However, in the new coordinate system that has been transformed via g_0, a similar equality does not hold in general,

$$\tilde{\mathbf{K}}_{Y_0} \neq \mathbf{B}\,\mathbf{K}_{Y_0}\,\mathbf{B}^T \tag{39}$$

Yet there is something that can be said if we consider the rotation and translation separately. First we will consider pure translation (*i.e.*, $R_0 = \mathbb{I}_3$);

$$\mathbf{Y} = \mathbf{X} + \tilde{\mathbf{t}}_0, \qquad \tilde{\mathbf{Y}} = \mathbf{B}\,\mathbf{Y} = \tilde{\mathbf{X}} + \mathbf{B}\,\tilde{\mathbf{t}}_0, \qquad \text{and} \qquad \mathbf{K}_{Y_0} = \mathbf{K}_{X_0} \tag{40}$$

Also,

$$\tilde{\mathbf{K}}_{Y_0} = \mathbf{B}\,\mathbf{K}_{Y_0}\,\mathbf{B}^T = \mathbf{B}\,\mathbf{K}_{X_0}\,\mathbf{B}^T \tag{41}$$

Next, consider pure rotation (*i.e.*, $\mathbf{t}_0 = \mathbf{0}$);

$$\mathbf{Y} = \tilde{\mathbf{R}}_0\,\mathbf{X} \qquad \text{and} \qquad \mathbf{K}_{Y_0} = \tilde{\mathbf{R}}_0\,\mathbf{K}_{X_0}\,\tilde{\mathbf{R}}_0^T \tag{42}$$

However

$$\tilde{\mathbf{Y}} = \begin{pmatrix} \tilde{\mathbf{R}}_0^{n-3}\,\mathbf{X}_{1,n-3} \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_{1,n-3} \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} \tag{43}$$

We can recover $\mathbf{Y}$ using $\mathbf{A} = \mathbb{I} \oplus \mathbb{I} \oplus \cdots \oplus \mathbb{I} \oplus R_0 \oplus R_0 \in \mathbb{R}^{3n\times 3n}$ via

$$\mathbf{Y} = \mathbf{A}\,\mathbf{B}^T \begin{pmatrix} \mathbf{Y}_{1,n-3} \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \mathbf{A}\,\mathbf{B}^T\,\tilde{\mathbf{Y}} \tag{44}$$

Then one can write

$$(\mathbf{Y} - \mathbf{Y}_0)^T\mathbf{K}_{Y_0}(\mathbf{Y} - \mathbf{Y}_0) = (\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}_0)^T\mathbf{B}\mathbf{A}^T\mathbf{K}_{Y_0}\mathbf{A}\mathbf{B}^T(\tilde{\mathbf{Y}} - \tilde{\mathbf{Y}}_0) \tag{45}$$

Therefore, for pure rotation

$$\tilde{\mathbf{K}}_{Y_0} = \mathbf{B}\mathbf{A}^T\mathbf{K}_{Y_0}\mathbf{A}\mathbf{B}^T \tag{46}$$

Using these two cases, we can generate any rigid body motion of one conformation relative to another. For example, if we want the coordinate frames attached to two conformations ($\mathbf{X}_0$ and $\mathbf{X}_0'$) to be separated by $g_0 = (R_0, \mathbf{t}_0)$, we could apply $g_{R_0} = (R_0, \mathbf{0})$ to $\mathbf{X}_0'$ and $g_{t_0} = (\mathbb{I}, -\mathbf{t}_0)$ to $\mathbf{X}_0$.