Modeling Loop Entropy

Gregory S. Chirikjian

Department of Mechanical Engineering
Johns Hopkins University
Baltimore, MD 21218, USA

Abstract

Proteins fold from a highly disordered state into a highly ordered one. Traditionally, the folding problem has been stated as one of predicting 'the' tertiary structure from sequential information. However, new evidence suggests that the ensemble of unfolded forms may not be as disordered as once believed, and that the native form of many proteins may not be described by a single conformation, but rather an ensemble of its own. Quantifying the relative disorder in the folded and unfolded ensembles as an entropy difference may therefore shed light on the folding process. One issue that clouds discussions of 'entropy' is that many different kinds of entropy can be defined: entropy associated with overall translational and rotational Brownian motion, configurational entropy, vibrational entropy, conformational entropy computed in internal or Cartesian coordinates (which can even be different from each other), conformational entropy computed on a lattice; each of the above with different solvation and solvent models; thermodynamic entropy measured experimentally, etc. The focus of this work is the conformational entropy of coil/loop regions in proteins. New mathematical modeling tools for the approximation of changes in conformational entropy during transition from unfolded to folded ensembles are introduced. In particular, models for computing lower and upper bounds on entropy for polymer models of polypeptide coils both with and without end constraints are presented. The methods reviewed here include kinematics (the mathematics of rigid-body motions), classical statistical mechanics and information theory.

Keywords: Protein folding, Entropy, Conformation, Ensemble, Convolution, Rigid-body Motion, Probability Density Function, Polymer, Information Theory

1 Introduction

In a classic observation, Anfinsen observed the spontaneous and repeatable folding of a protein from a highly disordered state into a highly ordered one [3]. From this result and others that followed, it has been inferred over the years that similar processes work for wide classes of proteins. But exactly how unstructured is the unfolded/denatured state ? And how structured is the native state ?

New evidence suggests that the ensemble of unfolded forms may not be as disordered as once believed and that the native form may not be as rigid as one might expect. In this light, protein folding is a transformation of a high-conformational-entropy ensemble into a lower one. But how high is high, and how low is low ? In order to answer such questions, some new mathematical and computer models will be helpful. Therefore, new mathematical tools for the approximation of conformational entropy in the unfolded and folded ensembles are introduced here. A number of related tools already exist in other fields. These are reviewed, adapted and developed further. In particular, lower and upper bounds on entropy are derived for polymer models of polypeptide chains, both with and without constraints on the positions and orientations of the ends. The methods reviewed here include kinematics (the mathematics of rigid-body motions as studied in the field of Robotics), information theory, and functional analysis on Lie groups (which, in part, considers how probability density functions of group-valued argument combine and propagate). In particular, we attach reference frames to polypeptide chains as shown in Figure 1, where the origin of the i^{th} frame is located at the i^{th} C_α atom with a unique orientation defined by the $C_\alpha - C'$ bond and the plane defined by the $C_\alpha - C' = O$ atoms as in [49]. We use the distributions of relative motion between consecutive residues to characterize backbone conformational entropy. Side-chain motions are computed relative to these reference frames and we show how to compute the associated side-chain entropy. These new and powerful methods make it possible to approximate changes in entropy between relatively ordered and disordered states without using traditional sampling techniques.

To summarize, the main contributions of this work are:

- A method for generating the distribution of relative positions and orientations of polymer-like polypeptide coils is presented, building on prior work in the Robotics literature.
- The distribution of end-constrained loop conformations is obtained from this information by applying Bayes' rule.
- Quantitative bounds on the associated loop entropy are derived, and the change in loop entropy resulting from constraining one end relative to the other is computed.

The computational complexity of this approach is low enough that it can be implemented on a single-processor personal computer running standardized software such as Matlab.

The remainder of this work is structured as follows. A comprehensive review of the literature is provided in Subsection 1.1. This is followed by a review of the necessary concepts from Statistical Mechanics in Subsection 1.2, and of background mathematics in Subsection 1.3. Section 2 applies these

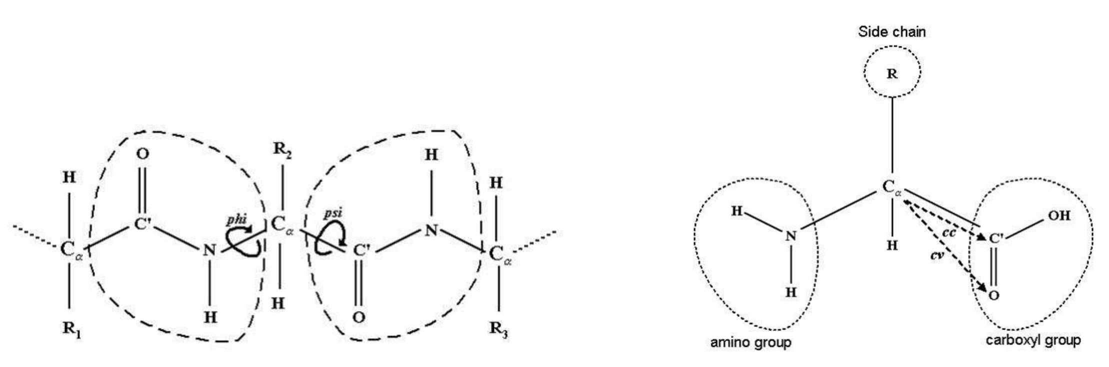


Figure 1: Reference Frames Attached to a Polypeptide Chain: (left) Dihedral angle definitions; (right) Attaching frames to C_α atoms in a canonical way.

techniques to compute lower and upper bounds on the entropy of the unfolded ensemble of polypeptide conformations. Section 3 develops bounds on the entropy of the folded ensemble. Section 4 demonstrates the methodology with some closed-form examples. Finally, Section 5 summarizes the results and maps out future directions.

1.1 Literature Review

Protein folding is often viewed graphically as a funnel from the polymer-like ensemble of unfolded states to the native state [10]. Changes in backbone entropy between unfolded and native states have been measured experimentally [22]. And NMR has been shown to be a useful tool for experimentally observing conformational fluctuations in proteins in general [52, 60, 83]. A growing body of literature suggests that ‘the native state’ of certain proteins may not be as ordered as once believed [28, 9, 78, 63, 29]. On the other hand, recent studies suggest that the unfolded ensemble is not as disordered as once believed [70], and that sequential interactions and sterics provide strong constraints on possible folding pathways [32, 36, 31, 61, 4]. Furthermore, conformational entropy of the native ensemble is believed to play an important role in binding [7, 34]. For these reasons, the development of analytical and computational models of entropy in protein loops with and without end constraints provides a way to compare the relative amount of disorder in the folded and unfolded cases.

Many statistical mechanical treatments of protein folding have been performed, e.g., [21, 80, 20, 26]. In some studies, full chemical detail is used in molecular dynamics [51, 40], yet it appears that this level of detail may not be required for successful prediction of folding [65]. Furthermore, when computing statistical quantities such as entropy, sufficient data must be obtained in high dimensional configuration or phase spaces in order to obtain robust results. This is almost impossible to do at a fully detailed level. Therefore, simplified statistical models such as those presented in this work may be useful. Furthermore, while the emphasis here is loop entropy in proteins, the methodology presented here can in principle be applied to RNA structures. Models of loop entropy in nucleic acids have been presented in [13, 53, 84].

The author's original field is the kinematic geometry of snakelike (or 'hyper-redundant') robot arms with many degrees of freedom [15, 16]. A tool which is useful for the analysis of all positions and orientations reachable by the 'gripper' at the distal end of this kind of arm is *noncommutative harmonic analysis* [18]. This mathematical tool combines ideas from group theory and Fourier analysis [71, 77, 58, 35, 72], and can be used to compute convolutions and diffusions of functions on Lie groups, such as the rotation group or rigid-body motion group [82, 18]. This is a particularly useful tool in the quantitative analysis of the distribution of all possible reachable gripper positions and orientations. Such quantities are quite similar to those encountered in polymer statistical mechanics. In polymer theory, distributions of relative end-to-end distance and orientation of backbone points and their tangents play central roles, as described in [6, 8, 24, 25, 27, 33, 37, 56, 68]. With the tool of noncommutative harmonic analysis, distributions in all six dimensions of rigid-body motion (three translational and three rotational) can be obtained, and marginals of these distributions can be taken to yield those which are commonly of interest in polymer physics (such as the distribution of end-to-end distances or relative orientations) [14]. This approach has been taken by the author in a series of papers, particularly concerned with semi-flexible polymers in which there is internal bending and torsional stiffness [17, 19]. The case of statistical distributions when semi-flexible polymers have internal joints and rigid bends has also been addressed using these methods [87, 88]. And it has been shown that this method can be applied to more general polymers including unfolded polypeptide chains [44]. Similar tools can be used to analyze large amounts of geometric data in the protein data bank [5] such as statistics of helix-helix crossing angle [50] and the relative pose (position and orientation) between alpha carbons in proteins [14, 49].

Of course, the author is not the only (and not even the first) member of the robotics community to attempt to transfer theoretical and computational tools from that field to study structural biology and biophysical phenomena. Lozano-Perez and coworkers have applied methods from robot motion planning and artificial intelligence to a number of problems in structural biology and rational drug design [66, 79]. Latombe and his students have applied methods from robot motion planning to explore configuration spaces and do energy minimization in the context of protein structures [54, 47, 38]. These build on the method of probabilistic roadmaps [41]. Amato and coworkers [75, 73, 1, 2] and Kavraki [74, 85, 23, 69] have been leaders in the application of robotics techniques in computational biology. The cyclic descent algorithm for robot kinematics has been applied to protein loops [11], as has other methods from kinematics [55, 42, 43]. It has been fashionable recently in engineering to consider proteins as examples of molecular machines [57, 45].

1.2 Statistical Mechanics

In classical equilibrium statistical mechanics, the Boltzmann distribution is defined as

$$f(\mathbf{p},\mathbf{q}) = \frac{1}{Z}\exp\left(-\beta\,\mathcal{H}(\mathbf{p},\mathbf{q})\right) \tag{1}$$

where the partition function is defined as

$$Z = \int_{\mathbf{q}}\int_{\mathbf{p}} \exp\left(-\beta\,\mathcal{H}(\mathbf{p},\mathbf{q})\right)\, d\mathbf{p}\, d\mathbf{q}. \tag{2}$$

Here $\beta = 1/k_BT$ (k_B is the Boltzmann constant and T is temperature measured in degrees Kelvin), $p_i = \mathbf{p}\cdot\mathbf{e}_i$ is the momentum conjugate to the i^{th} generalized coordinate $q_i = \mathbf{q}\cdot\mathbf{e}_i$, $\mathcal{H}$ is the Hamiltonian for the system, and $d\mathbf{p}\, d\mathbf{q} = dp_1 \cdots dp_N dq_1 \cdots dq_N$ for a system with N degrees of freedom. The range of integration is over all possible states of the system. The Boltzmann distribution describes the probability density of all states of a system at equilibrium.

The full set of generalized coordinates, $\{\mathbf{q}\}$, describes the *configuration* of the system, which includes overall rigid-body motion, and the intrinsic structural degrees of freedom. These intrinsic degrees of freedom can be further broken down into 'hard' degrees of freedom such as bond angles and bond lengths which do not vary substantially from referential values, and 'soft' degrees of freedom, such as torsion angles, that can vary widely. The hard degrees of freedom describe vibrational states and the soft degrees of freedom describe *conformational* changes, i.e., motions due to rotations around covalent chemical bonds. While the words 'configuration' and 'conformation' are often used interchangeably in the literature, the distinction between them as defined above is important in this work.

For any classical mechanical system the Hamiltonian is of the form

$$\mathcal{H}(\mathbf{p},\mathbf{q}) = \frac{1}{2}\mathbf{p}^T\{M^{-1}(\mathbf{q})\}\mathbf{p} + V(\mathbf{q}) \tag{3}$$

where $V(\mathbf{q})$ is the potential energy and $M(\mathbf{q})$ is the mass matrix (also called the mass metric tensor) [62].

The Gibbs formula for entropy of an ensemble described by $f(\mathbf{p},\mathbf{q})$ is

$$S = -k_B \int_{\mathbf{p}}\int_{\mathbf{q}} f(\mathbf{p},\mathbf{q}) \log f(\mathbf{p},\mathbf{q})\, d\mathbf{p}\, d\mathbf{q}. \tag{4}$$

Mathematically, 'continuous entropy' as defined above can take on negative values (and the entropy in the limiting case of a Dirac delta function goes to negative infinity). As explained in [12], this is very different than discrete entropy. Physically, continuum theory and classical mechanics break down at very small scales in phase space. By definition, a discretization of phase space is chosen such that $S = 0$ corresponds to the most ordered system that is physically possible, which is when all the states in an ensemble are contained in the same smallest possible element of discretized phase space. This is not the same as discretizing conformational space on a coarse lattice, as is often done in polymer simulations. The effects of discretization of continuous entropy are discussed in [12].

As a practical matter, there are several limitations in using (4) as a computational tool. First, there is some debate about what molecular potentials to use. On the one hand, the accuracy of ab initio potentials derived from first principles for small molecules and then applied to macromolecular simulations can be questioned. On the other hand, the accuracy of statistical potentials derived from structural data is limited by the richness of the databases from which they are extracted. For different perspectives on this debate see [30, 39, 46, 48, 59, 76]. Second, the number of degrees of freedom in macromolecules is so high (many thousands for a protein in continuum solvent, and perhaps millions when including explicit solvent degrees of freedom), that it is not possible to approximate $f(\mathbf{p}, \mathbf{q})$ with any degree of fidelity. (If the number of sample values required to accurately estimate a probability density function in one degree of freedom is K, then one would expect to need K^{2N} samples to approximate a probability density function (pdf) in a $2N$-dimensional phase space). If K is on the order of 10 to 100 and N ranges from thousands to millions, this is clearly intractable. One way to circumvent this problem is to compute only marginals of the full Boltzmann distribution, which as explained below, allows one to establish bounds on the true value of entropy.

Due to the structure of the Hamiltonian (3), it is easy to see that in general the Boltzmann distribution can not be separated into a product of configurational and momentum distributions, so $f(\mathbf{p}, \mathbf{q}) \neq f_p(\mathbf{p}) f_q(\mathbf{q})$ (due to the dependence of the mass matrix on configuration), and so the thermodynamic entropy is bounded by the entropies of each marginal as[1] $S \leq S_p + S_q$ where the configurational entropy is

$$S_q = -k_B \int_{\mathbf{q}} f_q(\mathbf{q}) \log f_q(\mathbf{q}) \, d\mathbf{q}. \tag{5}$$

and it is often assumed that S_p is constant. In fact, when the generalized coordinates are the Cartesian coordinates of the positions of all atoms in a macromolecule so that $\mathbf{q}$ becomes the $3n$-dimensional vector of all such positions, denoted here as $\mathbf{x} = [\mathbf{x}_1^T, ..., \mathbf{x}_n^T]^T$, then $f(\mathbf{p}, \mathbf{x}) = f_p(\mathbf{p}) f_x(\mathbf{x})$ and $S = S_p + S_x$.

[1] Using results from information theory [67, 12].

Furthermore, in this special case, the mass matrix is diagonal and constant, $M(\mathbf{x}) = M_0$, and

$$f_p(\mathbf{p}) = \frac{1}{Z_p} \exp\left(-\frac{1}{2}\mathbf{p}^T M_0^{-1} \mathbf{p}/k_B T\right)$$

and so

$$S_p = \log\{(2\pi e\, k_B T)^{3n/2} |M_0|^{\frac{1}{2}}\} \tag{6}$$

is in fact constant at constant temperature, without having to assume anything. It follows that $\Delta S = \Delta S_x$, which is not necessarily true for general choices of coordinates, including dihedral angles.

Under a change of coordinates $\mathbf{x} = \mathbf{x}(\mathbf{q})$ it is generally the case that $S_x \neq S_q$ because the computation of

$$S_x = -k_B \int_{\mathbf{x}} f_x(\mathbf{x}) \log f_x(\mathbf{x})\, d\mathbf{x} \tag{7}$$

in an alternative coordinate systems (such as dihedral angles) becomes

$$S_x = -k_B \int_{\mathbf{q}} f_x(\mathbf{x}(\mathbf{q})) \log f_x(\mathbf{x}(\mathbf{q}))\, |\det J(\mathbf{q})| d\mathbf{q}$$

which is not generally equal to S_q unless $|\det J(\mathbf{q})| = 1$. Therefore, when refering to configurational entropy, it is important to distinguish between Cartesian configurational entropy and dihedral configurational entropy unless $|\det J(\mathbf{q})| = 1$.

In some scenarios it is convenient to subdivide the configurational degrees of freedom into the categories: rigid-body, hard and soft, so that $\mathbf{q} = (\mathbf{q}_{rb}, \mathbf{q}_{hard}, \mathbf{q}_{soft})$. It can be shown that the determinants of the mass and Jacobian matrices for chain structures can be written as functions proportional to the form $w_1(\mathbf{q}_{rb}) \cdot w_2(\mathbf{q}_{hard})$. Similarly, it is a common modeling assumption that for a system not subjected to an external force field, and with sufficiently hard degrees of freedom, that

$$V(\mathbf{q}_{rb}, \mathbf{q}_{hard}, \mathbf{q}_{soft}) = V_1(\mathbf{q}_{hard}) + V_2(\mathbf{q}_{soft}).$$

Assumptions such as these lead to the separability of the partition function into a product, and the separability of entropy into a sum of terms:

$$S = S^{rb} + S^{hard} + S^{soft}.$$

Since the rigid-body term is the same for all ensembles of a given system in the same volume and temperature,

$$\Delta S = \Delta S^{hard} + \Delta S^{soft}.$$

We will focus on methods for computing Cartesian conformational entropy, ΔS_x, using the concept of convolution on the rigid-body-motion group. When the hard degrees of freedom are treated as rigid, $\Delta S^{hard} \rightarrow 0$, and $\Delta S_x \rightarrow \Delta S^{soft}$. In the remainder of this work we will examine S_x for: (a) polymer-like ensembles with rotatable bonds and free ends; and (b) polymer-like loop regions with end constraints.

1.3 Mathematics Review

When considering models of polypeptide chains, it often will be convenient to treat parts of the chain as rigid. For example, the plane of the peptide bond can be considered rigid, as can a cluster of side-chain atoms such a methyl group. At a coarser level, one might consider an alpha helix to be a rigid object. At a coarser level still, a whole domain might be approximated as a rigid body.

Therefore, it is clear that at various levels of detail, when characterizing the conformational entropy of a protein, it is conceivable that attaching reference frames to the rigid elements and recording the set of all possible rigid-body motions between these elements is a way to describe the conformational part of the Boltzmann distribution, and therefore get at the conformational entropy via Gibbs' formula.

In this section, a coordinate-free review of rigid-body motions is presented. More detailed reviews and comparisons of various parameterizations such as Euler angles, Cayley parameters, etc. can be found in [18].

1.3.1 Mathematics of Rigid-Body Motion

The group of rigid-body motions, which is also called the Special Euclidean group and is denoted $SE(3)$, is the semi direct product of $(\mathbb{R}^3, +)$ (three-dimensional Euclidean space endowed with the operation of vector addition) with the special orthogonal group, $SO(3)$, which consists of 3×3 rotation matrices together with the operation of matrix multiplication. In both instances, the word 'special' means that reflections are excluded and only physically allowable isometries of three-dimensional space are allowed.

We denote elements of $SE(3)$ as $g = (\mathbf{a}, A) \in SE(3)$ where $A \in SO(3)$ and $\mathbf{a} \in \mathbb{R}^3$. For any $g = (\mathbf{a}, A)$ and $h = (\mathbf{r}, R) \in SE(3)$, the group law is written as $g \circ h = (\mathbf{a} + A\mathbf{r}, AR)$, and $g^{-1} = (-A^T\mathbf{a}, A^T)$. Alternately, one may represent any element of $SE(3)$ as a 4×4 homogeneous transformation matrix of the form

$$g = \begin{pmatrix} A & \mathbf{a} \\ & \\ \mathbf{0}^T & 1 \end{pmatrix},$$

in which case the group law is matrix multiplication. The bottom row in these matrices, which consists of three zeros (i.e., $\mathbf{0}^T$ is the transposed, or row, vector corresponding to the column vector of zeros, $\mathbf{0}$) and the number one, is a placeholder which ensures that the matrix multiplication reproduces the correct group operation. In the above matrix, $A \in SO(3)$ denotes rotations and $\mathbf{a} \in \mathbb{R}^3$ denotes translations of a reference frame which when attached to a rigid body represent the motion of that body from the reference position and orientation defined by the identity element $e = (\mathbf{0}, I)$.

In Lie theory[2], the exponential mapping from the Lie algebra to a corresponding Lie group plays an important role [18]. In the current context, the Lie group of interest is $SE(3)$, and the corresponding Lie algebra is $se(3)$, which consists of all matrices formed by linear combinations of the following basis elements:

$$E_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad E_2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix};$$

$$E_3 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad E_4 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix};$$

$$E_5 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad E_6 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

For small (infinitesimal) motions around the identity (null motion), $g \approx I + X \in SE(3)$ where $X \in se(3)$. However, for larger motions this is not true. For those unfamiliar with his terminology, definitions and properties important to our formulation has been provided in the book [18]. The essential thing to know is that elements of $se(3)$ and $SE(3)$ can both be viewed as 4×4 matrices, however while it makes sense to add elements of $se(3)$ (i.e., velocities add), it only makes sense to multiply elements of

[2] Named after Norwegian mathematician Marius Sophus Lie (1842 - 1899).

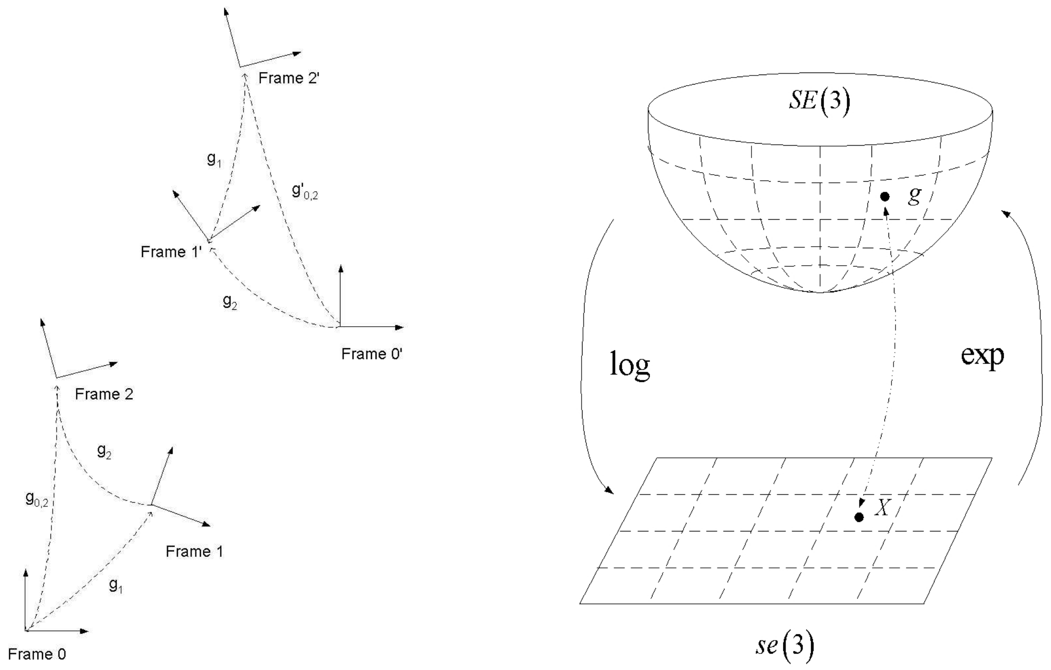


Figure 2: (left) Rigid-body Transformations between Reference Frames form a Noncommutative Lie Group ($g_1 \circ g_2 \neq g_2 \circ g_1$); (right) The Exponential Map

$SE(3)$. Furthermore, by the matrix exponential mapping, it is possible to produce elements of $SE(3)$ from those in $se(3)$, and vice versa using the matrix logarithm:

$$\exp : se(3) \rightarrow SE(3) \quad \text{and} \quad \log : SE(3) \rightarrow se(3).$$

Figure 2(left) illustrates that the composition of rigid-body motions is not a commutative operation. Figure 2(right) shows the relationship between the Lie algebra $se(3)$ consisting of infinitesimal motions (which form a linear vector space), and $SE(3)$ consisting of large motions (which form a curved manifold, which is a Lie group).

For small translational (rotational) displacements from the identity along (about) the i^{th} coordinate axis, the homogeneous transforms representing infinitesimal motions look like

$$\exp(\epsilon E_i) \approx I + \epsilon E_i \tag{8}$$

where I is the 4×4 identity matrix, $|\epsilon| << 1$, and $\exp(X) = I + X + X^2/2 + \cdots$ is the matrix exponential defined by the Taylor series of the usual exponential function evaluated with a matrix rather than a scalar. For example,

$$\exp(\theta E_3) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \exp(yE_5) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and for small values expanding $\sin\theta \approx \theta$ and $\cos\theta \approx 1$ it is easy to see that (8) holds for the example on the left. For the example on the right, (8) holds even for large values of y.

The 'exponential parametrization'

$$g = g(\chi_1, \chi_2, ..., \chi_6) = \exp\left(\sum_{i=1}^{6} \chi_i E_i\right) \tag{9}$$

is a useful way to describe relatively small rigid-body motions because, unlike the Euler angles, it does not have singularities near the identity.

One defines the 'vee' operator, $\vee$, such that for any

$$X = \sum_{i=1}^{6} \chi_i E_i,$$

$$X^{\vee} = \begin{pmatrix} \chi_1 \\ \chi_2 \\ \vdots \\ \chi_6 \end{pmatrix}.$$

The 6×6 adjoint matrix, Ad_g, is defined by the expression

$$Ad_g(X^{\vee}) = (gXg^{-1})^{\vee},$$

and explicitly if $g = (\mathbf{a}, A)$ then

$$Ad_g = \begin{pmatrix} A & 0 \\ \mathbf{a} \times A & A \end{pmatrix},$$

where $\mathbf{a} \times A$ denotes the matrix resulting from the cross product of $\mathbf{a}$ with each column of A.

The vector of exponential parameters, $\boldsymbol{\chi} \in \mathbb{R}^6$, can be obtained from $g \in G$ with the formula

$$\boldsymbol{\chi} = (\log g)^{\vee}. \tag{10}$$

The action of an element of the motion group, $g = (\mathbf{a}, A)$, on a vector $\mathbf{x}$ in three-dimensional space is defined as $g \cdot \mathbf{x} = A\mathbf{x} + \mathbf{a}$. In contrast, given a function $f(\mathbf{x})$, we can translate and rotate the function by g as $f(g^{-1} \cdot \mathbf{x}) = f(A^T(\mathbf{x} - \mathbf{a}))$. The fact that the inverse of the transformation applies under the function (rather than the transformation itself) in order to implement the desired motion is directly analogous to the case of translation on the real line. For example, given a function on the real line, $f(x)$, with its mode at $x = 0$, if we want to translate the whole function in the positive x direction by amount ξ so that the mode is at $x = \xi$, we compute $f(x - \xi)$ (not $f(x + \xi)$). This is a very important point to understand in order for the rest of this work to make sense. Figure 3(left) illustrates the shifting of a function under rigid-body motion geometrically.

1.3.2 Manipulations of Functions of Rigid-Body Motion

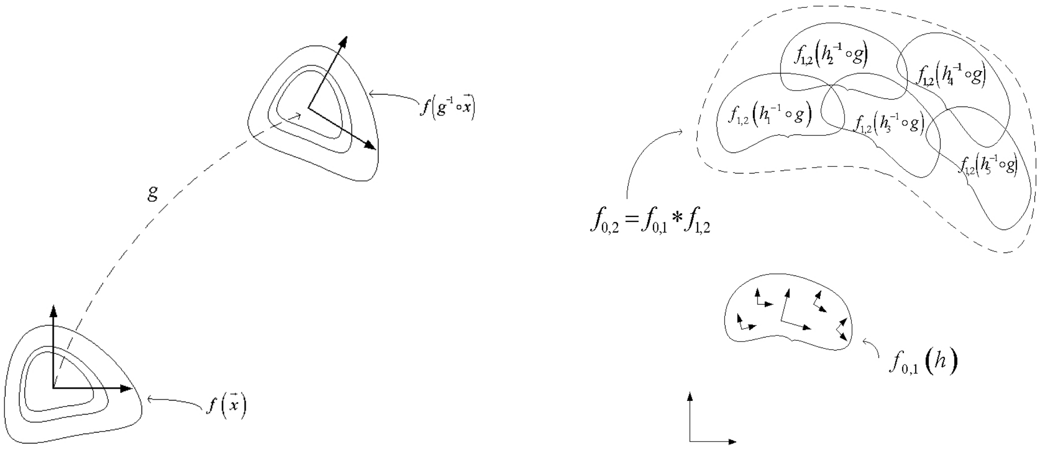
Suppose that three rigid bodies labeled 0, 1 and 2 are given, with reference frames attached to each, and assume that only sequentially adjacent bodies interact. Suppose also that body 0 is fixed in space and the ensemble of all possible motions of body 1 with respect to 0 are recorded, and motions of 2 with respect to 1 are also recorded. Then we have two functions of motion, $f_{0,1}(g)$ and $f_{1,2}(g)$ which together describe the conformational variability of this simple system. If we are interested in knowing the probability distribution describing the ensemble of all possible ways that body 2 can move relative to body 0, how is this obtained ? In fact, it is computed via the convolution on $SE(3)$ [18]:

$$f_{0,2}(g) = (f_{0,1} * f_{1,2})(g) = \int_G f_{0,1}(h) f_{1,2}(h^{-1} \circ g) dh \tag{11}$$

What this says is that the distribution $f_{1,2}(g)$ is shifted through all possible rigid-body motions, h, weighted by the frequency of occurrence of these motions, $f_{0,1}(h)$, and integrated over all values of $h \in G$ (G is just short for 'Group', which throughout this work is the group of rigid-body motions, $SE(3)$). Figure 3(right) illustrates this geometrically.

Explicitly, what is meant by this integral ? Let us assume for the moment that rotations are parameterized using Euler angles. The range of the Euler angles is $0 \leq \alpha, \gamma \leq 2\pi$ and $0 \leq \beta \leq \pi$. In this parametrization the volume element for G is given by

$$dg = \frac{1}{8\pi^2} \sin\beta d\alpha d\beta d\gamma dr_1 dr_2 dr_3,$$



$f(g^{-1}\circ\vec{x})$
g
$f(\vec{x})$
$f_{1,2}(h_2^{-1}\circ g)$
$f_{1,2}(h_4^{-1}\circ g)$
$f_{1,2}(h_1^{-1}\circ g)$
$f_{1,2}(h_3^{-1}\circ g)$
$f_{1,2}(h_5^{-1}\circ g)$
$f_{0,2}=f_{0,1}*f_{1,2}$
$f_{0,1}(h)$

Figure 3: (left) Action of a Motion on a Function; (right) Convolution of Functions of Rigid-Body Motion

which is the product of the volume elements for $\mathbb{R}^3$ ($d\mathbf{r} = dr_1 dr_2 dr_3$), and for $SO(3)$ ($dR = \frac{1}{8\pi^2}\sin\beta d\alpha d\beta d\gamma$). The normalization factor in the definition of dR is so that $\int_{SO(3)} dR = 1$. The volume element for $SE(3)$ can also be expressed in the exponential coordinates described in the previous subsection, in which case

$$dg = |J(\boldsymbol{\chi})| d\chi_1 \cdots d\chi_6$$

where $|J(\boldsymbol{\chi})|$ is a Jacobian determinant for this parametrization. The Jacobian can be computed using the formula

$$J(\boldsymbol{\chi}) = \left[\left(g^{-1}\frac{\partial g}{\partial \chi_1}\right)^{\vee}, \cdots, \left(g^{-1}\frac{\partial g}{\partial \chi_6}\right)^{\vee}\right]$$

and it can be shown that $|J(\mathbf{0})| = 1$ and so close to the identity the Jacobian factor in this parametrization can be ignored (which is not true for many other parameterizations, including the Euler angles).

The fact that the volume element is invariant to right and left translations, i.e,

$$dg = d(h \circ g) = d(g \circ h),$$

is well known in certain communities (See e.g. [77], [71]).

A convolution integral of the form in (11) can be written in the following equivalent ways:

$$(f_{0,1} * f_{1,2})(g) = \int_G f_{0,1}(z^{-1}) f_{1,2}(z \circ g) dz = \int_G f_{0,1}(g \circ k^{-1}) f_{1,2}(k) dk \tag{12}$$

where the substitutions $z = h^{-1}$ and $k = h^{-1} \circ g$ have been made, and the invariance of integration under shifts and inversions is used.

The concept of convolution on $SE(3)$ will be central in the formulation that follows.

One can define a Gaussian distribution on the six-dimensional Lie group $SE(3)$ much in the same way as is done on $\mathbb{R}^6$ provided that: (1) the covariances are small; and (2) the mean is located at the identity. The reason for these conditions is because near the identity, $SE(3)$ resembles $\mathbb{R}^6$ which means that $dg \approx d\chi_1 \cdots d\chi_6$ and we can define the Gaussian in the exponential parameters as

$$f(g(\boldsymbol{\chi})) = \frac{1}{(2\pi)^3|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}\boldsymbol{\chi}^T\Sigma^{-1}\boldsymbol{\chi}) \tag{13}$$

Given two such distributions that are shifted as $f_{i,i+1}(g_{i,i+1}^{-1} \circ g)$, each with 6×6 covariance $\Sigma_{i,i+1}$, then it can be shown that the mean and covariance of the convolution $f_{0,1}(g_{0,1}^{-1} \circ g) * f_{1,2}(g_{1,2}^{-1} \circ g)$ respectively will be of the form $g_{0,2} = g_{0,1} \circ g_{1,2}$ and [81]:

$$\Sigma_{0,2} = Ad_{g_{1,2}}^{-1} \Sigma_{0,1} Ad_{g_{1,2}}^{-T} + \Sigma_{1,2}. \tag{14}$$

This provides a method for computing covariances of two concatenated segments, and this formula can be iterated to compute covariances of chains without having to compute convolutions directly. This is demonstrated numerically in the context of robotic arms in [81].

2 Computing Bounds on the Entropy of the Unfolded Ensemble

2.1 End-to-End Position and Orientation Distributions and the Cartesian Conformational Entropy of Serial Polymer Chains

Consider a polymer consisting of a serial chain of $n+1$ essentially rigid monomer units numbered from 0 to n. Attach a frame of reference to the i^{th} such unit. Let g_i denote the rigid-body motion from the reference frame of the zeroth unit to that attached to the i^{th}. Let $g_{k,k+1}$ denote the relative motion from body k to body $k+1$. Then $g_i = g_{0,i} = g_{0,1} \circ g_{1,2} \circ \cdots \circ g_{i-1,i}$ will be the cumulative motion from body 0 to body i. The relationship between these reference frames is described in Figure 4.

In a purely pairwise energy model, only the interactions between adjacent units are important. In this simplest model, the probability of the relative pose $g_{i,i+1} = g_i^{-1} \circ g_{i+1}$ taking a particular value is given by

$$f_{i,i+1}(g_{i,i+1}) = (1/Z_{i,i+1}) \exp(-\beta V(g_{i,i+1})).$$

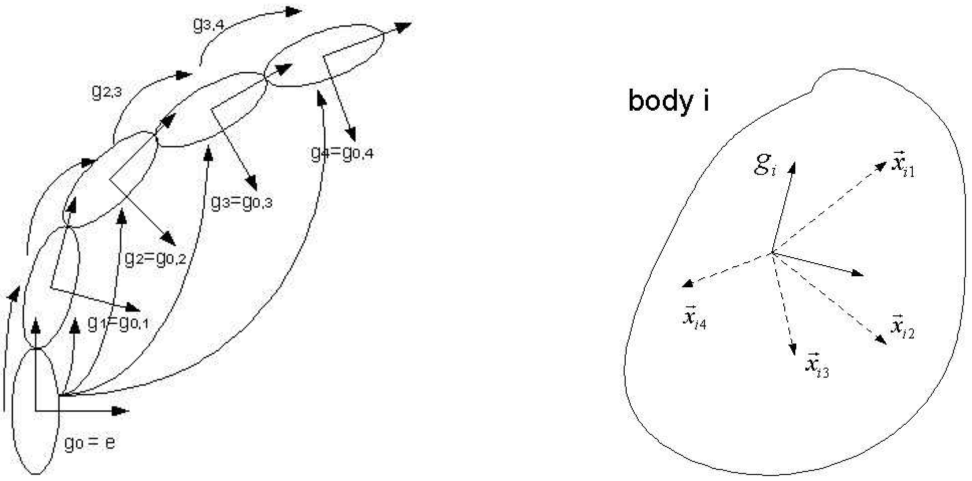


Figure 4: (left) Relative and Absolute Reference Frames attached to the Chain; (right) The relative positions of mass points within body i.

Then, the conformational distribution described in terms of rigid-body poses is

$$f(g_1, g_2, ..., g_n) = \prod_{i=0}^{n-1} f_{i,i+1}(g_i^{-1} \circ g_{i+1}) \tag{15}$$

where $g_0 = e$, the identity. This is related to the end-to-end position and orientation distribution

$$f_{0,n}(g_n) = (f_{0,1} * f_{1,2} * \cdots * f_{n-1,n})(g_n), \tag{16}$$

which is an n-fold convolution of the form in Section 1.3.2, by marginalization of (15) as

$$f_{0,n}(g_n) = \int_G \cdots \int_G f(g_1, g_2, ..., g_n) dg_1 \cdots dg_{n-1}.$$

This is illustrated in Figure 5.

Equation 15 represents a generalization of the classical polymer models in which only pairwise interactions are considered. If the frames of reference g_i and g_{i+1} are attached at the C_α atoms of residues i and $i+1$ in a polypeptide, then the function $f_{i,i+1}(g_{i,i+1})$ would be the six-dimensional generalization of a Ramachandran map that could include small bond angle bending, warping of the peptide plane and even bond stretching. If one chooses not to model these effects, then the classical Ramachandran map [64] can be reflected by appropriately defining $f_{i,i+1}(g_{i,i+1})$, as has been done in [44]. This is consistent with the Flory isolated pair model [33], which has been challenged in recent years [61]. However, as an upper bound on conformational entropy, it may still be useful in some contexts.

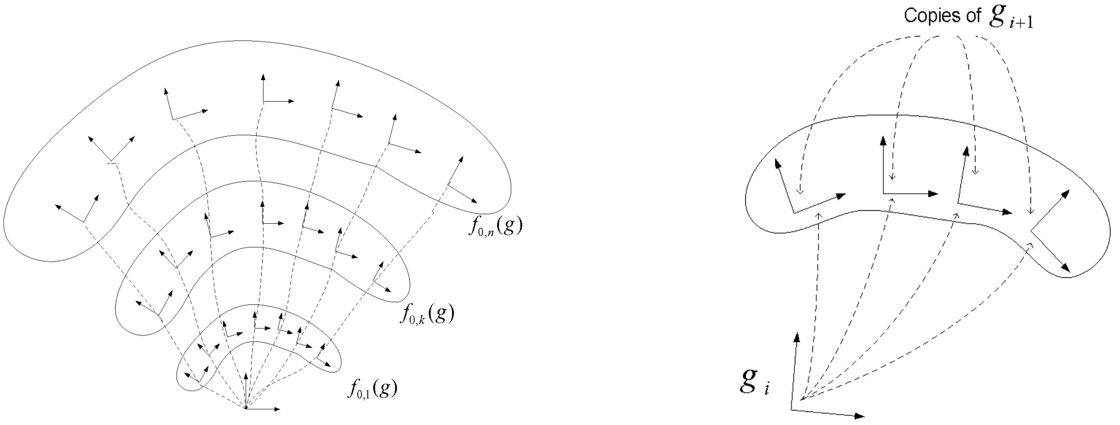


Figure 5: Kinematic Covariance Propagation: (left) In the absence of other constraints, distributions describing the allowable rigid-body motions between consecutive residues 'add' by convolution, resulting in a spreading out of probability density in position and orientation, $f_{0,i}(g_i)$, as i increases; (right) a zoomed in view of the probabilistic relationship between reference frames i and $i+1$ embodied by the functions $f_{i,i+1}(g_{i,i+1})$.

Note that since $g_n = (\mathbf{r}, R)$ describes both the end-to-end position and orientation of the distal end of the chain relative to the proximal end, we can marginalize further to obtain quantities such as the end-to-end distance distribution, or end-to-end orientational distribution. These quantities (or several of their moments) can be measured directly from a variety of experimental measurements.

In order to convert these probabilities into a form that is directly useful for computing Cartesian conformational entropy, we must know the positions of all atoms in each of the i rigid monomer units.

Given $f(g_1, g_2, ..., g_n)$ in (15) and given the family of probability density functions $\{\Delta_i(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_k})\}$ each of which describes the distribution of motions of the $i_k - i_1 + 1$ atoms within body i, it is possible to compute the full Cartesian conformational distribution as:

$$\rho(\mathbf{x}_1, ..., \mathbf{x}_N) = \int_G \cdots \int_G f(g_1, g_2, ..., g_n) \prod_{i=1}^{n} \Delta_i(g_i^{-1} \cdot \mathbf{x}_{i_1}, ..., g_i^{-1} \cdot \mathbf{x}_{i_k}) dg_1 \cdots dg_n \tag{17}$$

where $N = i_n$ is the total number of atoms in the chain and $\mathbf{x}_i = [\mathbf{x}_{i_1}^T, ..., \mathbf{x}_{i_k}^T]^T$ is the composite vector of Cartesian coordinates of all positions in the i^{th} body.

$\Delta_i(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_k})$ is a probability density on $3(i_k - i_1 + 1)$-dimensional Euclidean space. In other words,

$$\int_{\mathbb{R}^3} \cdots \int_{\mathbb{R}^3} \Delta_i(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_k}) d\mathbf{x}_{i_1} \cdots d\mathbf{x}_{i_k} = 1.$$

As an example, when the i^{th} body is modeled as being perfectly rigid,

$$\Delta_i(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_k}) = \prod_{j=i_1}^{i_k} \delta(\mathbf{x}_j - \mathbf{x}_j^0)$$

where $\mathbf{x}_j^0$ is the fixed position of atom j as seen in the frame of reference g_i affixed to rigid body i. In contrast, if body i is an articulated side chain, averaging over all of its conformational states would result in a Δ_i which is not a sum of Dirac delta functions.

In some cases it may be useful to compute the full pose entropy of the chain:

$$S_g = -\int_G \cdots \int_G f(g_1, g_2, ..., g_n) \log f(g_1, g_2, ..., g_n) dg_1 \cdots dg_n. \tag{18}$$

2.2 Modeling Excluded Volume Effects

The phantom polymer chain model in which the effects of excluded volume are ignored is clearly not a realistic model, but it can be used as a baseline onto which self-avoidance can be built. In a polypeptide, residue i interacts with residues $i+1$,..., $i+4$ substantially as well as more sequentially distant residues. These interactions are not only responsible for the formation of secondary structures, but also substantially winnow down the available conformational space [32]. Clearly this has implications for the entropy. More specifically, polymer models can be used to compute upper bounds on the conformational entropy in polypeptides. And these bounds can be made tighter by incorporating the effects of steric clash into modified versions of the conformational probability distributions.

To begin, let's compute the density of body i. This can either be done directly by, for example, averaging body i over all possible side-chain conformations. Or, it can be done by first computing each marginal of the density function Δ_i as:

$$d_{i_j}(\mathbf{x}_{i_j}) = \int_{\mathbf{x}_{i_1} \in \mathbb{R}^3} \cdots \int_{\mathbf{x}_{i_{j-1}} \in \mathbb{R}^3} \int_{\mathbf{x}_{i_{j+1}} \in \mathbb{R}^3} \cdots \int_{\mathbf{x}_{i_k} \in \mathbb{R}^3} \Delta_i(\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_k}) d\mathbf{x}_{i_1} \cdots \mathbf{x}_{i_{j-1}} \mathbf{x}_{i_{j+1}} \cdots d\mathbf{x}_{i_k}.$$

Then the average density of body i (normalized to be a probability density) is

$$d_i^{'}(\mathbf{x}) = \frac{1}{i_k - i_1 + 1} \sum_{i_j = i_1}^{i_k} d_{i_j}(\mathbf{x}).$$

The overlap of bodies in the chain is illustrated in Figure 6.

Therefore, if body i is moved by rigid body motion g_i, and likewise for body j, we can compute an

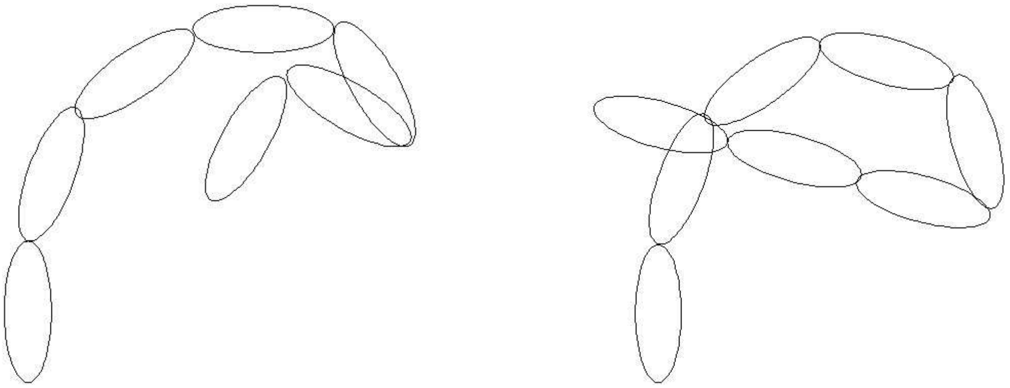


Figure 6: Conformations to be removed from the Phantom Chain Ensemble: (left) Local Overlaps; (right) Nonlocal Overlaps.

estimate of their overlap (averaged over all deformations of the bodies) as

$$w_{ij}(g_i, g_j) = \int_{\mathbb{R}^3} d_i^{'}(g_i^{-1} \cdot \mathbf{x}) d_j^{'}(g_j^{-1} \cdot \mathbf{x}) d\mathbf{x}.$$

A general property of integration over all of three-dimensional space is that it is invariant under rigid-body motions. Therefore if, we make the change of variables $\mathbf{y} = g_i^{-1} \circ \mathbf{x}$, then we find that

$$w_{ij}(g_i, g_j) = w_{ij}(e, g_i^{-1} \circ g_j) = w_{ij}(g_j^{-1} \circ g_i, e).$$

Clearly when the two bodies do not overlap, $w_{ij} = 0$. Otherwise they will have some positive value. One can imagine evaluating w_{ij} as the argument of a 'sigmoid function', which sharply ramps up from zero to one, where it then plateaus at higher values. The resulting $W_{ij}(g_i^{-1} \circ g_j) = 1 - \exp(-(w_{ij}(g_i, g_j))^2/2\sigma^2)$ (for some small value of σ) would effectively window out all values of the rigid-body motions g_i and g_j that contribute to nonphysical overlaps. Then the original $f(g_1, g_2, ..., g_n)$ in (15) could be replaced with one of the form

$$f_{ex}(g_1, g_2, ..., g_n) = Cf(g_1, g_2, ..., g_n) \prod_{i<j}^{n} (1 - W_{ij}(g_i^{-1} \circ g_j)) \tag{19}$$

where C is the normalization required to make f_{ex} a probability density function. Note that the product in this expression is not only over sequentially local pairs of bodies, but rather all bodies, where the '$i < j$' simply avoids double counting. In this way, a phantom polymer model that generates $f(g_1, g_2, ..., g_n)$ can be viewed as the starting point for a more realistic model that includes steric constraints.

2.3 Bounding Cartesian Conformational Entropy

Practically speaking computing such high-dimensional integrals as in (17) or (19) can impose a computational problem, except when simple closed-form expressions such as Gaussians are used. If we seek an upper bound on Cartesian conformational entropy, marginals can be computed and information-theoretic bounds can be employed. Performing such marginalization, one finds

$$\rho_i(\mathbf{x}_i) = \int_G f_{0,i}(g_i)\Delta_i(g_i^{-1} \cdot \mathbf{x}_{i_1}, ..., g_i^{-1} \cdot \mathbf{x}_{i_k})dg_i \tag{20}$$

In the case when one representative point is chosen per residue (for example, the C_α atom, which is where the reference frame for the residue is usually attached), we have $k = 1$. Then $i = i_1$, and since $\mathbf{x}_i^0 = \mathbf{0}$ due to the way the reference frame is attached, we can write

$$\rho_i(\mathbf{x}_i) = \int_G f_{0,i}(g)\delta(g^{-1} \cdot \mathbf{x}_i)dg.$$

If $g = (\mathbf{r}, R)$, then $\delta(g^{-1} \cdot \mathbf{x}_i) = \delta(R^T(\mathbf{x}_i - \mathbf{r})) = \delta(\mathbf{x}_i - \mathbf{r})$ and so we can get the positional distribution of the i^{th} C_α atom by marginalizing the full pose distribution over orientations as

$$\rho_i(\mathbf{x}_i) = \int_{SO(3)} \int_{\mathbb{R}^3} f_{0,i}(\mathbf{r}, R)\delta(\mathbf{x}_i - \mathbf{r})d\mathbf{r}\, dR = \int_{SO(3)} f_{0,i}(\mathbf{x}_i, R)dR. \tag{21}$$

The conformational entropy of the backbone represented by C_α atoms is then bounded from below by the entropy of individual marginals (with the tightest lower bound resulting from the maximum of these). The loop entropy will be bounded from above by the sum of entropies from all of the marginals. Therefore,

$$\max_i S_i \leq S_x \leq \sum_{i=1}^{n} S_i \quad \text{where} \quad S_i = -\int_{\mathbb{R}^3} \rho_i(\mathbf{x}_i) \log \rho_i(\mathbf{x}_i) d\mathbf{x}_i \tag{22}$$

and $\mathbf{x} = [\mathbf{x}_1^T, \mathbf{x}_2^T, ..., \mathbf{x}_n^T]^T$.

3 Approximating Entropy of the Loops in the Folded Ensemble

The native ensemble of a protein is characterized by a relatively high degree of order. However, the native form is not completely rigid. In particular, loop/coil regions connecting secondary structures can exhibit large motions. Here we model the ends of these loops as being fixed at specific positions and orientations, as illustrated in Figure 7. Bounds on the contribution of loop motions to overall entropy

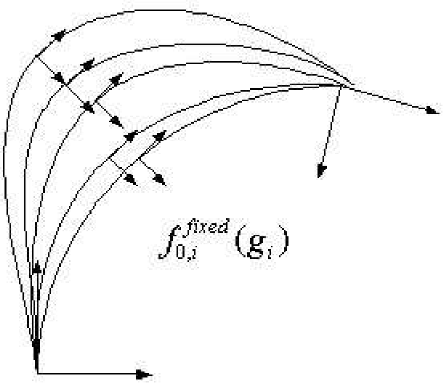


Figure 7: Using Density Information to Determine Probabilities of Conformations that Obey End Constraints

are discussed here.

If $f(g_1, ..., g_n)$ is the conformational distribution function describing the positions and orientations of all bodies in the system with respect to the proximal end of the chain, then if we fix the distal end at a specific pose, g_{end}, the resulting distribution will the the conditional density

$$f^{fix}(g_1, g_2, ..., g_{n-1}; g_{end}) = f(g_1, ..., g_n | g_n = g_{end}) = f(g_1, ..., g_{n-1}, g_{end})/f_{0,n}(g_{end}). \tag{23}$$

The entropy of this distribution in some cases can either be computed directly, or each of the marginals can be computed as

$$f_{0,i}^{fix}(g_i; g_{end}) = f_{0,i}(g_i) f_{i,n}(g_i^{-1} \circ g_{end})/f_{0,n}(g_{end}). \tag{24}$$

The reason for this is that in the definition of $f(g_1, ..., g_n)$, the variable g_i appears in only the two multiplied terms: $f_{i-1,i}(g_{i-1}^{-1} \circ g_i) \cdot f_{i,i+1}(g_i^{-1} \circ g_{i+1})$. Marginalizaing over g_1 through g_{i-1} results in $f_{0,i}(g_i)$. If g_i were the identity element, then marginalizing over g_{i+1} through g_{n-1} would yield $f_{i,n}(g_{end})$. However, since in general $g_i \neq e$, this result is shifted by g_i to yield $f_{i,n}(g_i^{-1} \circ g_{end})$. Division by $f_{0,n}(g_{end})$ is the normalization required to make the result a pdf (since integration of the numerator in (24) over g_i is a convolution). This denominator is carried along from (23), which is a statement of Bayes' rule.

Intuitively, the entropy of a chain with fixed ends must be smaller than that of a chain with freely moving ends. This can be quantified when using (23) and (24), as will be shown in the examples in the next section.

4 Examples

In this section examples are used to illustrate the formulation presented earlier in this work. In both of these examples, a piece of flexible loop/coil connects relatively rigid structures. In the first example, the loop is considered to be a long phantom chain, whereas in the second it is considered to be a semi-flexible polymer. The reduction in conformational entropy associated with constraining the ends in both cases is examined.

4.1 Model 1: Long Loops Modeled as Gaussian Chains

Perhaps the most common model for the distribution of end-vector distribution in polymer theory is the Gaussian distribution:

$$f(g) = W(\mathbf{r}) = \left(\frac{3}{2\pi\langle r^2\rangle}\right)^{\frac{3}{2}} \exp\left[-\frac{3r^2}{2\langle r^2\rangle}\right] = \frac{1}{(2\pi)^{3/2}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}\mathbf{r}^T\Sigma^{-1}\mathbf{r}) \tag{25}$$

where $g = (\mathbf{r}, R) \in SE(3)$ and the chain is so flexible that the orientational part of the distribution is constant, and $\Sigma = (\langle r^2\rangle/3)I$ where I is the 3×3 identity matrix. This distribution is spherically symmetric (and hence depends only on $r = |\mathbf{r}|$). It is normalized so that it is a probability density function,

$$\int_{\mathbb{R}^3} W(\mathbf{r})dV = 4\pi\int_0^\infty W(\mathbf{r})r^2dr = 1,$$

satisfying

$$\int_{\mathbb{R}^3} W(\mathbf{r})|\mathbf{r}|^2dV = 4\pi\int_0^\infty W(\mathbf{r})r^4dr = \langle r^2\rangle.$$

The probability density function for a freely-jointed chain with n links, each of length l can be approximated as a Gaussian random walk with

$$\langle r^2\rangle = nl^2.$$

If we denote $W_n(\mathbf{r})$ to be the function for n links, then it is clear that in this simple model $W_{n_1} * W_{n_2} = W_{n_1+n_2}$. Therefore, in this simple model the conformational entropy is bounded as

$$\frac{3}{2}\log(2\pi enl^2/3) \leq S_r \leq \frac{3}{2}\sum_{k=1}^{n}\log(2\pi ekl^2/3)$$

using (22) where r takes the place of x.

The conformational entropy of the phantom chain that gives rise to this distribution is bounded from below by the entropy of the probability density function of the location of the terminal end, and from above by the sum of probabilities for each link from base to end, since these are marginals of the total conformational distribution $f(g_1, ..., g_n)$, which for this case is a function $W(\mathbf{r}_1, ..., \mathbf{r}_n) = \prod_{i=0}^{n-1} W_1(\mathbf{r}_{i+1} - \mathbf{r}_i)$ with $\mathbf{r}_0 = \mathbf{0}$ where W_1 is the effective one-bond Gaussian distribution with covariance matrix $\Sigma_1 = (l^2/3)I$. Therefore, it is possible to compute the Cartesian conformational entropy in this model exactly in closed form as

$$S_r = -\int_{\mathbf{r}_1} \cdots \int_{\mathbf{r}_n} W(\mathbf{r}_1, ..., \mathbf{r}_n) \log W(\mathbf{r}_1, ..., \mathbf{r}_n) d\mathbf{r}_1 \cdots d\mathbf{r}_n.$$

Since the chain is assumed to be uniform and the product of Gaussians is a Gaussian, we can use the fact that there is a closed-form formula for the entropy of a Gaussian in terms of its covariance matrix, together with the fact that in this example

$$\Sigma^{-1} = \frac{3}{l^2} \begin{pmatrix} 2I & -I & 0 & 0 & \cdots & 0 \\ -I & 2I & -I & 0 & \cdots & \vdots \\ 0 & -I & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & -I & 2I & -I \\ 0 & \cdots & 0 & 0 & -I & I \end{pmatrix}. \tag{26}$$

In principle, the entropy of a Gaussian chain with end position constraints can be computed using (23) and (24). In practice, there are some details that need to be addressed, which are addressed in Section 4.3.

Whereas here a Gaussian chain in which orientations diffuse rapidly was considered, the opposite extreme of a stiff chain is considered in the following subsection.

4.2 Model 2: Short Loops Modeled as Semi-flexible Polymers

Suppose that we have a semi-flexible loop, i.e., one that has local resistance to bending and twisting that reflects sequentially local steric constraints. Then each g_i will deviate only a relatively small amount from a constant reference pose, h_i, and so we write $g_i = h_i \circ \exp \boldsymbol{\chi}_i$ where $\|\boldsymbol{\chi}_i\| < 1$. This sort of assumption is consistent with findings in the literature. For example, [86] validated the use of semi-flexible polymer models to describe protein loop motions. The relative motion between adjacent reference frames will be

$$g_i^{-1} \circ g_{i+1} = \exp -\boldsymbol{\chi}_i \circ h_i^{-1} \circ h_{i+1} \circ \exp \boldsymbol{\chi}_{i+1}.$$

If the probability density $f_{i,i+1}$ is a Gaussian with mean at $h_i^{-1} \circ h_{i+1}$, then it will be of the form

$$f_{i,i+1}(g) = F_{\Sigma_{i,i+1}}((h_i^{-1} \circ h_{i+1})^{-1} \circ g)$$

where

$$F_{\Sigma_{i,i+1}}(\exp \boldsymbol{\chi}) = \frac{1}{(2\pi)^3 |\Sigma_{i,i+1}|^{\frac{1}{2}}} \exp(-\frac{1}{2} \boldsymbol{\chi}^T \Sigma_{i,i+1}^{-1} \boldsymbol{\chi}).$$

Therefore,

$$f_{i,i+1}(g_i^{-1} \circ g_{i+1}) = F_{\Sigma_{i,i+1}}((h_i^{-1} \circ h_{i+1})^{-1} \circ \exp -\boldsymbol{\chi}_i \circ h_i^{-1} \circ h_{i+1} \circ \exp \boldsymbol{\chi}_{i+1}).$$

For small motions between adjacent bodies, the approximation

$$[\log((h_i^{-1} \circ h_{i+1})^{-1} \circ \exp -\boldsymbol{\chi}_i \circ h_i^{-1} \circ h_{i+1} \circ \exp \boldsymbol{\chi}_{i+1})]^{\vee} \approx \boldsymbol{\chi}_{i+1} - Ad^{-1}_{h_i^{-1} \circ h_{i+1}} \boldsymbol{\chi}_i$$

has been proven to be accurate [81]. If as shorthand we define $A_{i,i+1} = Ad_{h_i^{-1} \circ h_{i+1}}$, then

$$f_{i,i+1}(g_i^{-1} \circ g_{i+1}) =$$

$$\frac{1}{(2\pi)^3 |\Sigma_{i,i+1}|^{\frac{1}{2}}} \exp -\frac{1}{2} [\boldsymbol{\chi}_i^T, \boldsymbol{\chi}_{i+1}^T] \begin{pmatrix} A_{i,i+1}^{-T} \Sigma_{i,i+1}^{-1} A_{i,i+1}^{-1} & -A_{i,i+1}^{-T} \Sigma_{i,i+1}^{-1} \\ -\Sigma_{i,i+1}^{-1} A_{i,i+1}^{-1} & \Sigma_{i,i+1}^{-1} \end{pmatrix} \begin{bmatrix} \boldsymbol{\chi}_i \\ \boldsymbol{\chi}_{i+1} \end{bmatrix}. \tag{27}$$

This, together with the product in (15), leads $f(g_1, g_2, ..., g_n)$ to be a Gaussian distribution in the variable $\boldsymbol{\chi} = [\boldsymbol{\chi}_1^T, ..., \boldsymbol{\chi}_n^T]^T$ with an inverse covariance of the form

$$\Sigma^{-1} = \tag{28}$$

$$\begin{pmatrix}
\Sigma_{0,1}^{-1}+\Sigma_{1,2}'^{-1} & -A_{1,2}^{-T}\Sigma_{1,2}^{-1} & 0 & \cdots & 0 & 0 & 0 \\
-\Sigma_{1,2}^{-1}A_{1,2}^{-1} & \Sigma_{1,2}^{-1}+\Sigma_{2,3}'^{-1} & -A_{2,3}^{-T}\Sigma_{2,3}^{-1} & 0 & \ddots & \ddots & 0 \\
0 & -\Sigma_{2,3}^{-1}A_{2,3}^{-1} & \ddots & \ddots & \ddots & 0 & 0 \\
\vdots & 0 & \ddots & \ddots & \ddots & 0 & \vdots \\
0 & \ddots & \ddots & -\Sigma_{n-3,n-2}^{-1}A_{n-3,n-2}^{-1} & \Sigma_{n-2,n-1}^{-1}+\Sigma_{n-3,n-2}'^{-1} & -A_{n-2,n-1}^{-T}\Sigma_{n-2,n-1}^{-1} & 0 \\
0 & \ddots & 0 & 0 & -\Sigma_{n-2,n-1}^{-1}A_{n-2,n-1}^{-1} & \Sigma_{n-2,n-1}^{-1}+\Sigma_{n-1,n}'^{-1} & -A_{n-1,n}^{-T}\Sigma_{n-1,n}^{-1} \\
0 & 0 & 0 & \cdots & 0 & 0 & \Sigma_{n-1,n}^{-1}
\end{pmatrix}.$$

where $\Sigma'_{i,i+1} = A_{i,i+1}\Sigma_{i,i+1}A_{i,i+1}^T$.

The entropy S_g for the case with free ends is then given by the formula (18), which is relatively efficient to compute due to the block tri-diagonal form of Σ^{-1}.

Obtaining the entropy for the case when both ends are fixed in position and orientation is also possible within this model. Using the fact that the adjoint is a homomorphism, i.e., $Ad(g_1 \circ g_2) = Ad(g_1)Ad(g_2)$ and $Ad(g^{-1}) = Ad^{-1}(g)$, this generalizes to the concatenation of n reference frames that vary around values h_i as

$$\Sigma_{0*1*\cdots*n} = \sum_{k=0}^{n} Ad_{h_{k,n}}^{-1}\,\Sigma_k\,Ad_{h_{k,n}}^{-T} \tag{29}$$

In order to compute S_g for the case when the distal end of the chain is fixed at $g_n = g_{end}$, we would use (23) with the covariance of $f_{0,n}(g)$ being given by (29). Conditioning of Gaussians by Gaussians yields Gaussians, the entropy of which can be computed in closed form in principle. However, there are some subtle issues that need to be addressed, as discussed below.

4.3 From Covariance Matrices to Entropy

It is one thing to have bounds such as (22). It is another to have a closed-form expression for the actual quantity of interest. Here (26) and (28) are used to compute entropy. This follows from the fact that for a $d(n)$-dimensional Gaussian distribution with covariance Σ, the entropy is given as [67, 12]

$$S = \log\{(2\pi e)^{d(n)/2}|\Sigma|^{\frac{1}{2}}\}. \tag{30}$$

Here we use the notation $d(n)$ to denote the dimension of the covariance matrix, which is $d(n) = 3n$ for positional Gaussian, and $d(n) = 6n$ for the semi-flexible case. In other words, we can write $d(n) = d_0 \cdot n$.

We will consider the case when entropy change is due to fixing the ends. Consider a chain (either Gaussian or semi-flexible), and let $\mathbf{x}_1$,...,$\mathbf{x}_n$ denote the variables describing the kinematic state of the n segments (i.e., $\mathbf{x}_i = \mathbf{r}_i \in \mathbb{R}^3$ for the Gaussian chain and $\mathbf{x}_i = \boldsymbol{\chi}_i \in \mathbb{R}^6$ for the semi-flexible chain). Let us denote $\mathbf{x} = [\mathbf{x}_1^T, ..., \mathbf{x}_{n-1}^T]^T$, $\mathbf{y} = \mathbf{x}_n \in \mathbb{R}^{d_0}$ and $\mathbf{z} = [\mathbf{x}^T, \mathbf{y}^T]^T$. Then $f(\mathbf{x}_1, ..., \mathbf{x}_n)$ can be written as

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{d_0 n/2} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z}\right).$$

The entropy of this distribution is computed simply as (30). However, the end-constrained case is somewhat more involved.

The conditional probability density describing the ensemble of end-constrained conformations is of the form

$$f(\mathbf{x}|\mathbf{y}) = f(\mathbf{x}, \mathbf{y})/f(\mathbf{y})$$

where $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{z})$ (with $\mathbf{y}$ held fixed rather than being a variable) and the marginal distribution $f(\mathbf{y})$ is given by

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{d_0/2} |\Sigma_{yy}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}^T \Sigma_{yy}^{-1} \mathbf{y}\right)$$

where

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}.$$

The conditional distribution will then be

$$f(\mathbf{x}|\mathbf{y}) = \frac{1}{(2\pi)^{d_0(n-1)/2} |\Lambda|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}[\mathbf{x} - \mathbf{x}_0]^T \Lambda^{-1} [\mathbf{x} - \mathbf{x}_0]\right) \tag{31}$$

where

$$\Lambda = \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} = [(\Sigma^{-1})_{xx}]^{-1}$$

and

$$\mathbf{x}_0 = \Sigma_{xy} \Sigma_{yy}^{-1} \mathbf{y}.$$

In principle we now have everything we need to compute entropy differences. However, in practice, there is an implicit assumption about polymer distribution functions that must be addressed. Namely, even though the chain length is L, and hence the distal end cannot reach outside a ball or radius L

centered at the proximal end, for the sake of convenience we will accept distributions with infinitely long tails. Another way to say this is that as long as the pdf decays rapidly enough and all integrals over a ball of radius L centered at the origin can be replaced by integrals over an infinitely large ball, then things will work out fine. Such calculations include computing probabilities and entropies from probability densities. In other words, we have simplifications such as

$$\int_{-L}^{L} e^{-x^2} dx \approx \int_{-\infty}^{\infty} e^{-x^2} dx.$$

While this is perfectly reasonable when the Gaussians are centered at the origin, it will no longer be the case when we shift them by significant amounts. In other words, even though the value of an integral over an infinite range is invariant under shifts, this is not the case for integration over finite intervals:

$$\int_{-L}^{L} e^{-(x-L/2)^2} dx \neq \int_{-L}^{L} e^{-x^2} dx.$$

This is important in the context of the current discussion because the conditional pdf in (31) is shifted from the origin by a vector $\mathbf{x}_0$. In other words, if we fix the distal end of the chain at an arbitrary $\mathbf{y}$, then this distribution of interest in $d_0 \cdot (n-1)$-dimensional space will not be centered at the origin, and the infinite integral used to approximate integration over a ball of radius $L = nl$ centered at the origin that resulted in the normalization constant $[(2\pi)^{d_0(n-1)/2}|\Lambda|^{\frac{1}{2}}]^{-1}$ will no longer be a valid approximation. The computation of this constant, and the computation of entropy then become a problem when $\|\mathbf{y}\|$ (and hence $\|\mathbf{x}_0\|$) is not very small relative to total chain length, $L = nl$. However, when it is very small, the integral can still be approximated as being over infinite-dimensional space because the overwhelming majority of the mass under the pdf will still be contained in the finite ball of radius L.

4.3.1 Entropy for the Gaussian Chain

For the Gaussian chain effectively the end constraint $\mathbf{r}_n = \mathbf{0}$ means that the chain forms a closed loop because the vectors $\{\mathbf{r}_i\}$ in this case are absolute positions of the i^{th} residue with respect to the proximal end.

The entropy difference between two ensembles described by Gaussians with dimensions $d(n)$ and $d(n-1)$ in the unconstrained and end-constrained states, respectively, will be

$$\Delta S = S_2 - S_1 = \log\{(2\pi e)^{d(n)/2}|\Sigma_2|^{\frac{1}{2}}\} - \log\{(2\pi e)^{d(n-1)/2}|\Sigma_1|^{\frac{1}{2}}\} = \log\left[(2\pi e)^{d_0/2}|\Sigma_2|^{\frac{1}{2}}/|\Sigma_1|^{\frac{1}{2}}\right]$$

$$= \frac{1}{2} \log \left[(2\pi e)^{d_0} |\Sigma_1^{-1}| / |\Sigma_2^{-1}| \right] . \tag{32}$$

The last equality means that there is no need to invert the matrices in (26) and (28) when computing entropy differences between the ensembles with free and fixed ends. This is useful, because in practice one usually is interested only in entropy differences, and the determinants of block-tridiagonal matrices can be computed very efficiently (in $O(n)$ computations for a chain of length n), whereas computing their inverses followed by taking the determinant can be an $O(n^3)$ operation.

4.3.2 Entropy of a Semi-flexible Chain

The entropy being considered is that defined in (18). For the semi-flexible chain the set $\{\boldsymbol{\chi}_i\}$ describes the relative small rigid-body displacements of the i^{th} residue with respect to a referential configuration. Therefore in this case the same tools developed in this work can be used to describe the entropy differences between the free and end-constrained cases for a somewhat different scenario than the Gaussian chain model. Namely, we can compute multiple reference conformations and consider small deviations around each. The reduction in entropy due to constraining both ends of the chain is then due to eliminating motions around multiple reference conformations (each with free ends) and only allowing motions around the one reference conformation that satisfies the required end conditions. This discussion is quantified below.

Imagine sampling the relative poses between adjacent amino acids in a loop at their K most populated isolated peaks. For example, K might be equal to 3 if we sample at the centers of the α, β and Ω regions of the ϕ-ψ plane. If each of these peaks are isolated, and the distributions around them are modeled as $SE(3)$ Gaussian distributions with small covariances and essentially non-overlapping tails, then the entropy associated with each of these conformational ensembles will be given by (30) where Σ is defined by (28). If the loop has n residues, then each reference conformation for $i = 1, ..., K^n$ will have its own entropy, S_i defined by these equations. If the relative weights of each of these reference conformations are given by $w_1, ..., w_{K^n}$, and if each conformation is disjoint, and there is minimal overlap between the associated conformational distributions around each, then the total entropy in the case of free ends can be approximated as

$$S_{free} \approx - \sum_{i=1}^{K^n} w_i \log w_i + \sum_{i=1}^{K^n} w_i S_i .$$

The entropy for the case when the distal end is fixed can be approximated by adding contributions from the subset of baseline conformations that approximately satisfy the end constraints.

5 Conclusions

This work reviewed and built on techniques from the fields of robotics, information theory and theoretical polymer science and applied these to model conformational entropy in protein loops. At the core of this presentation was the mathematics of rigid-body motion and associated statistical computations, as well as the use of inequalities from information theory for developing a rigorous mathematical treatment of the entropy of unfolded, partially folded and fully folded proteins. Models of conformational statistics in these three kinds of ensembles were reviewed and developed. These models were then applied to compute entropy differences. The various concepts of entropy in statistical mechanics, computational polymer science and information theory were reviewed. The distinction between conformational entropy computed in internal and Cartesian coordinates was made. Inequalities to bound the value of entropy from below and above were presented in cases when exact computations were judged to be intractable.

Acknowledgements

This work was performed with support from the National Institutes for Health under grant R01GM075310 and R01GM075310-04S1. The author would like to thank Profs. G. Rose and E. Lattman for their valuable comments, and Dr. W. Park for proofreading, and Dr. S. Lee and Ms. Y. Wang for creating some of the figures..

References

[1] Amato, N.M., Song, G., "Using motion planning to study protein folding pathways," *Journal of Computational Biology*, 9(2): 149-168, 2002.

[2] Amato, N.M., Dill, K.A., Song, G., "Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures," *Journal of Computational Biology*, 10(3-4):239-255 2003.

[3] Anfinsen, C.B., "Principles that govern folding of protein chains," *Science* 181(4096): 223-230, July 1973.

[4] Baldwin, R.L., Rose, G.D., "Is protein folding hierarchic? I. Local structure and peptide folding," *Trends in Biochemical Sciences*, 24(1):26-33, Jan 1999.

[5] 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 Research*, 28(1):235-242, Jan 1 2000.

[6] Birshtein, T.M., Ptitsyn, O.B., *Conformations of Macromolecules,* Interscience, New York, 1966.

[7] Boehr, D.D., Nussinov, R., Wright, P.E., "The role of dynamic conformational ensembles in biomolecular recognition," *Nature Chemical Biology* 5:789 - 796, 2009.

[8] Boyd, R.H., Phillips, P.J., *The Science of Polymer Molecules*, Cambridge Solid State Science Series, Cambridge University Press, Cambridge, 1993.

[9] Bracken, C., Iakoucheva, L.M., Rorner, P.R., Dunker, A.K., "Combining prediction, computation and experiment for the characterization of protein disorder," *Current Opinion in Structural Biology*, 14 (5): 570-576, Oct 2004.

[10] Bryngelson, J.D., Onuchic, J.H., Socci, N.D., Wolynes, P.G.,"Funnels, Pathways, and the Energy Landscape of Protein-Folding - A Synthesis," *Proteins-Structure,Function and Genetics*, 21(3):167-195, March 1995.

[11] Canutescu, A.A,, Dunbrack, R.L., "Cyclic coordinate descent: A robotics algorithm for protein loop closure," *Protein Science*, 12(5):963-972, May 2003.

[12] Chirikjian, G.S., *Stochastic Models, Information Theory, and Lie Groups*, Birkhäuser, 2009.

[13] Chirikjian, G.S., "Group Theory and Biomolecular Conformation, I.: Mathematical and computational models," *J. Phys.: Condens. Matter* 22, 323103 (2010).

[14] Chirikjian, G.S., "Conformational Statistics of Macromolecules Using Generalized Convolution," *Computational and Theoretical Polymer Science*, 11:143-153, Feb 2001.

[15] Chirikjian, G.S., Burdick, J.W., "A Modal Approach to Hyper-Redundant Manipulator Kinematics." *IEEE Transactions on Robotics and Automation* 10:343-354, 1994.

[16] Chirikjian, G.S., Burdick, J.W., "A Geometric Approach to Hyper-Redundant Manipulator Obstacle Avoidance," *ASME Journal of Mechanical Design*, 114:580-585, December 1992.

[17] Chirikjian, G.S., Kyatkin, A.B.,"An Operational Calculus for the Euclidean Motion Group with Applications in Robotics and Polymer Science," *J. Fourier Analysis and Applications*, 6(6):583-606, Dec 2000.

[18] Chirikjian, G.S., Kyatkin, A.B., *Engineering Applications of Noncommutative Harmonic Analysis*, CRC Press, Boca Raton, FL 2001.

[19] Chirikjian, G.S., Wang, Y., "Conformational Statistics of Stiff Macromolecules as Solutions to PDEs on the Rotation and Motion Groups," *Physical Review E*, 62(1):880-892, July 2000.

[20] Crippen, G.M., " A Gaussian statistical mechanical model for the equilibrium thermodynamics of barnase folding," *Journal of Molecular Biology.* 306(3):565-573, Feb 23 2001.

[21] Crippen, G.M., " Statistical mechanics of protein folding by cluster distance geometry," *Biopolymers*, 75(3):278-289, Oct 15 2004.

[22] D'Aquino, J.A., Gomez, J., Hilser, V.J., Lee, K.H., Amzel, L.M., Fieire, E., "The Magnitude of the Backbone Conformational Entropy Change in Protein Folding," *PROTEINS: Structure, Function, and Genetics*, 25:143-156, 1996.

[23] Das, P., Moll. M., Stamati, H., Kavraki, L. E., Clementi, C., "Low-dimensional Free-energy Landscapes of Protein-Folding Reactions by Nonlinear Dimensionality Reduction," *PNAS*, 103(26):98859890, June 17 2006.

[24] de Gennes, P.G., *Scaling Concepts in Polymer Physics*, Cornell University Press, 1979.

[25] des Cloizeaux, J., Jannink, G., *Polymers in Solution: Their Modelling and Structure*, Clarendon Press, Oxford, 1990.

[26] Dill, K.A., Fiebig, K.M., Chan, H.S., "Cooperativity in Protein-Folding Kinetics," *PNAS*, 90(5):1942-1946, March 1 1993.

[27] Doi, M., Edwards, S.F., *The Theory of Polymer Dynamics*, Clarendon Press, Oxford, 1986.

[28] Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., Uversky, V.N., "Flexible nets - The roles of intrinsic disorder in protein interaction networks," *FEBS Journal*, 272(20):5129-5148, Oct 2005.

[29] Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.R., Hipps, K.W., Ausio, J., Nissen, M.S., Reeves, R., Kang, C.H., Kissinger, C.R., Bailey, R.W., Griswold, M.D., Chiu, M., Garner, E.C., Obradovic, Z., "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling* 19 (1): 26-59 2001.

[30] Fang, Q.J., Shortle, D., "A consistent set of statistical potentials for quantifying local side-chain and backbone interactions," *Proteins-Structure, Function and Bioinformatics*, 60(1):90-96, July 1 2005.

[31] Fitzkee, N.C., Rose, G.D., "Reassessing random-coil statistics in unfolded proteins," *PNAS*, 101(34):12497-12502, August 24 2004.

[32] Fitzkee, N.C., Rose, G.D., "Sterics and solvation winnow accessible conformational space for unfolded proteins," *Journal of Molecular Biology*, 353(4):873-887, Nov 4 2005.

[33] Flory, P.J., *Statistical Mechanics of Chain Molecules*, John Wiley & Sons, 1969 (reprinted Hanser Publishers, Munich 1989).

[34] Frederick, K.K., Marlow, M.S., Valentine, K.G., Wand, A.J., "Conformational entropy in molecular recognition by proteins," *Nature*, 448:325-330, July 19, 2007.

[35] Gel'fand, I.M., Minlos, R.A., Shapiro, Z. Ya., *Representations of the Rotation and Lorentz Groups and Their Applications*, Pergamon Press, New York, 1963.

[36] Gong, H.P., Rose, G.D., "Does secondary structure determine tertiary structure in proteins?," *Proteins-Structure,Function and Bioinformatics*, 61(2):338-343, Nov 1 2005.

[37] Grosberg, A. Yu., Khokhlov, A.R., *Statistical Physics of Macromolecules*, American Institute of Physics, New York, 1994.

[38] Hsu, D., Latombe, J.C., Motwani, R., "Path planning in expansive configuration spaces," *International Journal of Computational Geometry and Applications*, 9(4-5):495-512, Aug-Oct 1999.

[39] Jernigan, R.L., Bahar, I., "Structure-derived potentials and protein simulations," *Current Opinion in Structural Biology*, 6(2):195-209, April 1996.

[40] Karplus, M., Weaver, D.L., "Protein-Folding Dynamics," *Nature*, 260(5550):404-406, 1976.

[41] Kavraki, L.E., Svestka, P., Latombe, J.C., Overmars, M.H., "Probabilistic roadmaps for path planning in high-dimensional configuration spaces," *IEEE Transactions on Robotics and Automation*, 12(4):566-580, Aug 1996.

[42] Kazerounian, K., Latif, K., Rodriguez, K., Alvarado, C., "Nano-kinematics for analysis of protein molecules," *Journal of Mechanical Design*, 127(4):699-711, July 2005.

[43] Kazerounian, K., "From mechanisms and robotics to protein conformation and drug design," *Journal of Mechanical Design*, 126(1):40-45, Jan 2004.

[44] Kim, J.S., Chirikjian, G.S., "A unified approach to conformational statistics of classical polymer and polypeptide models," *Polymer*, 46(25):11904-11917, Nov 28 2005.

[45] Kim, M.K., Jernigan, R.L., Chirikjian, G.S., "Rigid-cluster models of conformational transitions in macromolecular machines and assemblies," *Biophysical Journal*, 89(1):43-55, July 2005.

[46] Kortemme, T., Morozov, A.V., Baker, D., "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes," *Journal of Molecular Biology*, 326(4):1239-1259, Feb 28 2003.

[47] Lavalle, S.M., Finn, P.W., Kavraki, L.E., Latombe, J.C., "A randomized kinematics-based approach to pharmacophore-constrained conformational search and database screening," *Journal of Computational Chemistry*, 21(9):731-747, July 15 2000.

[48] Lazaridis, T., Karplus, M., " Effective energy function for proteins in solution," *Proteins-Structure, Function and Genetics*, 35(2):133-152, May 1 1999.

[49] Lee, S., Chirikjian, G.S., "Pose analysis of alpha-carbons in proteins," *International Journal of Robotics Research*, 24(2-3):183-210, Feb-March 2005.

[50] Lee, S., Chirikjian, G.S., "Inter-Helical Angle and Distance Preferences in Globular Proteins," *Biophysical Journal*, 86:1105-1117, Feb 2004.

[51] Levitt, M., "Protein Folding by Restrained Energy Minimization and Molecular-Dynamics," *Journal of Molecular Biology*, 170(3):723-764, 1983.

[52] Li, Z., Raychaudhuri, S., Wand, J., "Insights into the local residual entropy of proteins provided by NMR relaxation," *Protein Science* 5:2647-2650 (1996).

[53] Liu, L., Chen, S.-J., "Computing the conformational entropy for RNA folds," *J. Chem. Phys.* 132, 235104 (2010).

[54] Lotan, I., Schwarzer, F., Halperin, D., Latombe, J.C., "Algorithm and data structures for efficient energy maintenance during Monte Carlo simulation of proteins," *Journal of Computational Biology*, 11(5):902-932, Oct 2004.

[55] Manocha, D., Zhu, Y.S., Wright, W., "Conformational-Analysis of Molecular Chains Using Nano-Kinematics," *Computer Applications in the Biosciences*, 11(1):71-86, Feb 1995.

[56] Mattice, W.L., Suter, U.W., *Conformational Theory of Large Molecules, The Rotational Isomeric State Model in Macromolecular Systems*, Wiley, New York, 1994.

[57] Mavroidis, C., Dubey, A., Yarmush, M.L., "Molecular machines," *Annual Review of Biomedical Engineering*, 6:363-395, 2004.

[58] Miller, W. Jr., *Lie Theory and Special Functions*, Academic Press, New York, 1968; also see Miller, W. Jr., "Some Applications of the Representation Theory of the Euclidean Group in Three-Space," *Commun. Pure App. Math.*, 17:527-540, 1964.

[59] Moult, J., "Comparison of database potentials and molecular mechanics force fields," *Current Opinion in Structural Biology*, 7(2):194-199, 1997.

[60] Palmer, A.G. III, "Probing molecular motions by NMR," *Current Opinion in Structural Biology*, 7:732-737 (1997).

[61] Pappu, R.V., Srinivasan, R., Rose, G.D., "The Flory isolated-pair hypothesis is not valid for polypeptide chains: Implications for protein folding," *PNAS*, 97(23):12565-12570, Nov 7 2000.

[62] Patriciu, A., Chirikjian, G.S., Pappu, R.V., "Analysis of the conformational dependence of mass-metric tensor determinants in serial polymers with constraints," *Journal of Chemical Physics*, 121(24):12708-12720, Dec. 22 2004.

[63] Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D., Dunker, A.K., " Protein flexibility and intrinsic disorder," *Protein Science* 13 (1): 71-80 JAN 2004.

[64] Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V., "Stereochemistry of Polypeptide Chain Configurations," *Journal of Molecular Biology*, 7(1):95-99, 1963.

[65] Rhee, Y.M., Pande, V.S., "On the role of chemical detail in simulating protein folding kinetics," *Chemical Physics*, 323:66-77, 2006.

[66] Rienstra, C.M., Tucker-Kellogg, L., Jaroniec, C.P., Hohwy, M., Reif, B., McMahon, M.T., Tidor, B., Lozano-Perez, T., Griffin, R.G., "De novo determination of peptide structure with solid-state magic-angle spinning NMR spectroscopy," *PNAS*, 99(16):10260-10265, Aug 6 2002.

[67] Shannon, C.E., "A mathematical theory of communication," *Bell System Technical Journal*, Vol. 27, pp. 379-423 and pp. 623-656, July and October 1948.

[68] Skliros, A., Chirikjian, G.S., "Positional and Orientational Distributions for Locally Self-Avoiding Random Walks with Obstacles," *Polymer*, 49(6): 1701-1715, March 2008.

[69] Shehu, A., Clementi, C., Kavraki, L. E., "Modeling Protein Conformational Ensembles: From Missing Loops to Equilibrium Fluctuations," *Proteins: Structure, Function, and Bioinformatics*, 65(1):164-179, 2006.

[70] Shortle, D., Ackerman, M.S., "Persistence of native-like topology in a denatured protein in 8 M urea," *Science*, 293(5529): 487-489, July 20 2001.

[71] Sugiura, M., *Unitary Representations and Harmonic Analysis*, 2^{nd} edition, Elsevier Science Publisher, The Netherlands, 1990.

[72] Talman, J., *Special Functions*, W. A. Benjamin, Inc., Amsterdam, 1968.

[73] Tang, X.Y., Kirkpatrick, B., Thomas, S., Song, G., Amato, N.M., " Using motion planning to study RNA folding kinetics," *Journal of Computational Biology*, 12(6): 862-881, July 2005.

[74] Teodoro, M., Phillips, G. N. Jr., Kavraki, L. E.,"Molecular Docking: A Problem with Thousands of Degrees of Freedom," In *Proc. of the 2001 IEEE International Conference on Robotics and Automation (ICRA 2001)*, pp. 960966, IEEE press, Seoul, Korea, May 2001.

[75] Thomas, S., Song, G., Amato, N.M., "Protein folding by motion planning," *Physical Biology*, 2(4):S148-S155, Dec 2005.

[76] Vajda, S., Sippl, M., Novotny, J., "Empirical potentials and functions for protein folding and binding," *Current Opinion in Structural Biology*, 7(2):222-228, 1997.

[77] Vilenkin, N.J., Klimyk, A.U., *Representation of Lie Group and Special Functions*, Vol. 1-3, Kluwer Academic Publishers, The Netherlands, 1991.

[78] Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L.M., Cortese, M.S., Lawson, J.D., Brown, C.J., Sikes, J.G., Newton, C.D., Dunker, A.K., "DisProt: a database of protein disorder," *Bioinformatics* 21 (1): 137-140 JAN 1 2005.

[79] Wang, C.S.E., Lozano-Perez, T., Tidor, B., "AmbiPack: A systematic algorithm for packing of macromolecular structures with ambiguous distance constraints," *Proteins-Structure Function and Genetics*, 32(1):26-42, July 1 1998.

[80] Wang, J.Y., Crippen, G.M., " Statistical mechanics of protein folding with separable energy functions," *Biopolymers*, 74(3):214-220, June 15 2004.

[81] Wang, Y., Chirikjian, G.S., “Nonparametric Second-Order Theory of Error Propagation on the Euclidean Group,” *International Journal of Robotics Research*, 27(1112): 12581273, November/December 2008.

[82] Wang, Y., Chirikjian, G.S., “Workspace Generation of Hyper-Redundant Manipulators as a Diffusion Process on SE(N),” *IEEE Transactions on Robotics and Automation*, 20(3):399-408, June 2004.

[83] Yang, D., Kay, L.E., “Contributions to Conformational Entropy Arising from Bond Vector Fluctuations Measured from NMR-Derived Order Parameters: Application to Protein Folding,” *Journal of Molecular Biology*, 263:369-382 (1996).

[84] 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.* 128, 125107 (2008).

[85] Zhang, M., White, R. A., Wang, L., Goldman, R., Kavraki, L. E., Hassett, B.,“Improving Conformational Searches by Geometric Screening,” *Bioinformatics*, 21(5):624630, 2005

[86] Zhou, H.-X., “Loops in Proteins Can Be Modeled as Worm-Like Chains,” *J. Phys. Chem. B*, 105, 6763-6766 (2001).

[87] Zhou, Y., Chirikjian, G.S., “Conformational Statistics of Semi-Flexible Macromolecular Chains with Internal Joints,” *Macromolecules*, 39(5):1950-1960, 2006.

[88] Zhou, Y., Chirikjian, G.S., “Conformational statistics of bent semiflexible polymers,” *Journal of Chemical Physics*, 119(9):4962-4970, Sept. 1 2003.