Stochastic Modeling of Rotational Degrees of Freedom in Molecules

HARSHINDER SINGH

Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505
Department of Statistics, West Virginia University, Morgantown, WV 26506
his6@cdc.gov, hsingh@stat.wvu.edu

EUGENE DEMCHUK

Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505
School of Pharmacy, West Virginia University, Morgantown, WV 26506
eed5@cdc.gov

VLADIMIR HNIZDO

Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505
vbh5@cdc.gov

DAN S. SHARP

Health Effects Laboratory Division, National Institute for Occupational Safety and Health, Morgantown, WV 26505
zre6@cdc.gov

Abstract. The dynamics of molecular conformation depends primarily on random fluctuations in the dihedral angle space of the molecule. Investigation of these random fluctuations is important in studying the function of peptides, proteins and DNA molecules. It is also important in drug design and investigation of effects of toxins. Thus, the stochastic modeling of random fluctuations in the dihedral angle space is relevant to many areas in biomolecular sciences. These fluctuations can be large or small depending on the nature of the molecule and on ambient conditions, such as temperature and pressure. There can be multiple peaks in marginal distributions of the angles because of hindered rotations. The stochastic modeling of the joint distribution of these angles thus poses an interesting challenge. The challenge is even greater for modeling of dihedral angles of biological macromolecules such as proteins since these molecules have a very large number of rotational degrees of freedom, many of which are interdependent. We review a probabilistic approach to stochastic modeling of multidimensional rotational motions in dihedral angles of molecular systems.

1 Introduction

Understanding the structure, function and stability of biomolecules, such as peptides and proteins, bears on investigation of random fluctuations that occur in their atomic coordinates. Entropy is a cumulative measure of these stochastic movements, and it is one of the most important thermodynamic functions of molecular systems. Restriction or loosening of molecular motions has major implications on the molecular properties. It may cause conformational rearrangements such as folding or unfolding in proteins, affect macromolecular association, and even redirect ligand binding to another target receptor, a problem quite relevant to designing new drug therapies.

Bond lengths, bond angles and torsional (also called rotational or dihedral) angles determine internal coordinates of the molecule. Bond lengths and bond angles are relatively rigid coordinates compared to torsional angles, and thus random fluctuations in torsional angles make the major contribution to the internal entropy. The probabilistic modeling and theoretical analysis of multivariate circular distributions that are observed in stochastic processes dependent on torsional degrees of freedom is an important problem in biomolecular sciences.

In Section 2 of this paper, we describe statistical thermodynamic concepts for the benefit of the reader with no background in statistical thermodynamics. In Section 3, we review the stochastic modeling techniques for modeling random movements in the torsional angles of molecules, along with procedures for evaluation of internal entropy.

2 Statistical Concepts in Thermodynamics

Let $X_1, X_2, \ldots, X_m$ be the m atomic coordinate positions of a molecular system; these coordinates are assumed to be random variables. Let $f(x_1, x_2, \ldots, x_m)$ denote the probability density function of the coordinates. The entropy S associated with these coordinates is defined by

$$\begin{aligned} S &= k_B E[-\ln f(X_1, X_2, \ldots, X_m)] \\ &= -k_B \int \ldots \int \ln f(x_1, x_2, \ldots, x_m)\ f(x_1, x_2, \ldots, x_m) dx_1 dx_2 \ldots dx_m \quad (1) \end{aligned}$$

where k_B is Boltzmann's constant.

Without loss of generality, consider a situation when the potential and kinetic energy terms are independent. Let $V(x_1, x_2, \ldots, x_m)$ denote the potential energy associated with the system when $X_i = x_i$, $i = 1, 2, \ldots, m$. In statistical thermodynamics, the standard way of choosing the probability distribution of $X_1, X_2, \ldots, X_m$ is by using the maximum entropy principle. The probability distribution with probability density function $f(x_1, x_2, \ldots, x_m)$, which maximizes the entropy under the condition of a fixed mean potential energy E_0, i.e.,

$$\int \cdots \int V(x_1, x_2, \ldots, x_m) f(x_1, x_2, \ldots, x_m) dx_1 dx_2 \ldots dx_m = E_0,$$

is given by

$$f(x_1, x_2, \ldots, x_m) = \frac{1}{Z(\beta)} \exp\{-\beta V(x_1, x_2, \ldots, x_m)\} \tag{2}$$

where

$$Z(\beta) = \int \cdots \int \exp\{-\beta V(x_1, x_2, \ldots, x_m)\} dx_1 dx_2 \ldots dx_m \tag{3}$$

is the reciprocal of the normalizing constant and is called the configurational partition function. The variable β is taken as $1/k_B T$, where T is the temperature. The probability density function (2) is called the Boltzmann-Gibbs distribution of the system. It is easily seen that

$$E\{V(X_1, X_2, \ldots, X_m)\} = -\frac{d}{d\beta} \ln Z(\beta)$$

and

$$\mathrm{Var}\{V(X_1, X_2, \ldots, X_m)\} = -\frac{d^2}{d\beta^2} \ln Z(\beta).$$

The entropy associated with the probability density function f can be expressed as

$$S = k_B[\ln Z(\beta) + \beta E\{V(X_1, X_2, \ldots, X_m)\}].$$

The free energy of the system is defined by

$$\begin{aligned} A &= -\frac{1}{\beta} \ln Z(\beta) \\ &= E\{V(X_1, X_2, \ldots, X_m)\} - \frac{1}{\beta k_B} S. \end{aligned} \tag{4}$$

Thus, the Boltzmann-Gibbs distribution can be viewed as the probability distribution which minimizes the free energy of the system. More details on statistical thermodynamics are given, for example, in Landau and Lifshitz [1980].

3 Stochastic Modeling of Rotational Degrees of Freedom in Molecules

To a large extent, the random fluctuations in torsional angles (rotational degrees of freedom) determine the internal entropy associated with a system of bonded atoms — the molecule. Let $\Theta_1, \Theta_2, \ldots, \Theta_m$ be the m torsional angles of a molecule. These coordinates have been modeled in the literature assuming a multivariate normal distribution with the variance/co-variance matrix Σ [Karplus and Kushik, 1981; Levy *et al*, 1984]. In this model, the entropy associated with the angles is given by

$$S = \frac{mk_B}{2} + \frac{k_B}{2}\ln[(2\pi)^m|\Sigma|]. \tag{5}$$

In this case, the Boltzmann-Gibbs distribution is the multivariate normal distribution; and thus the potential energy of the system is a quadratic function of the coordinates.

This approach has been used to obtain configurational entropy of butane and decaglycine [Karplus and Kushik, 1981]. Although the multivariate normal distribution provides a way to model a large number of dependent coordinates in macromolecules, the assumption of a multivariate normal distribution does not hold in general. The multivariate normal distribution is not adequate for a molecule that has large fluctuations around rotatable bonds. Demchuk and Singh [2001] introduced a circular probability approach to modeling of torsional angles in molecules. As a case study, they modeled the torsional angle of the methanol molecule by a 3-mode von Mises distribution with the probability density function

$$f(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa\cos[3(\theta-\theta_0)]}, \ -\pi \le \theta < \pi, \tag{6}$$

where $\kappa > 0$ is the concentration parameter, $I_0(\kappa)$ is a modified Bessel function of order 0 and $\theta_0 = -2\pi/3$. The entropy of the von Mises distribution is given by

$$S = k_B\left[\ln 2\pi + \ln[I_0(\kappa)] - \kappa\frac{I_1(\kappa)}{I_0(\kappa)}\right]. \tag{7}$$

where $I_1(\kappa)$ is the modified Bessel function of order 1. Using the von Mises distribution as the Boltzmann-Gibbs distribution, the potential energy associated with the torsional angle of methanol is given by $V = (V_0/2)[1 - \cos[3(\theta - \theta_0)]$, where V_0 is the maximum torsional energy. The authors derived the distribution of torsional energy; it is bathtub-shaped with a probability density function given by

$$g(v) = \frac{1}{\pi I_0(\kappa)} \exp[\kappa(1 - \frac{2v}{V_0})] v^{-1/2} \ (V_0 - v)^{-1/2} \ , \ 0 \leq v \leq V_0. \tag{8}$$

The 3-mode von Mises distribution and the bathtub distribution provided excellent fits respectively to the molecular dynamic simulated data of torsional angle and of torsional energy for methanol.

In order to model two dependent torsional angles in a molecule, Singh *et al* [2002] proposed a bivariate circular distribution which is a natural torus analogue of the bivariate normal distribution. Let Θ_1 and Θ_2 be two circular random variables in the range of $[-\pi, \pi)$. The proposed joint probability density function for Θ_1 and Θ_2 is given by

$$f(\theta_1, \theta_2) = Ce^{\kappa_1 \cos(\theta_1 - \mu_1) + \kappa_2 \cos(\theta_2 - \mu_2) + \lambda \sin(\theta_1 - \mu_1) \sin(\theta_2 - \mu_2)}, \ -\pi \leq \theta_1, \theta_2 < \pi, \tag{9}$$

where κ_1, $\kappa_2 \geq 0$, $-\infty < \lambda < \infty$, $-\pi \leq \mu_1, \mu_2 < \pi$, and C is the normalization constant such that $f(\theta_1, \theta_2)$ is a pdf. The conditional distributions of this model are von Mises, and the marginals are symmetric and are either unimodal or bimodal depending on the configuration of parameters. The distribution has a natural generalization for allowing multiple modes in the marginal distributions. The authors applied this formalism to modeling two dependent torsional angles in a pentapeptide.

In general, the marginal distributions of torsional angles may adopt irregular shapes. Expanding the potential energy function in a Fourier series, Singh *et al* [2002] and Hnizdo *et al* [2002] approached situations in which the distributions of a torsional angle are of arbitrary shape. Let Θ denote such an angle and let $V(\theta)$ denote the effective torsional potential energy of the molecule when the torsional angle takes the value θ. The Boltzmann-Gibbs probability density function of Θ is given by

$$f(\theta) = \frac{1}{Z(\beta)} e^{-\beta V(\theta)}. \tag{10}$$

The authors considered the Fourier series expansion of $V(\theta)$,

$$V(\theta) = \sum_{j=0}^{\infty} [a_j \cos(j\theta) + b_j \sin(j\theta)] \tag{11}$$

where a_j and b_j are fixed constants. Taking a finite number of terms into consideration, the proposed probability density function $f(\theta)$ of Θ is given by

$$f(\theta) = \frac{1}{2\pi I_0(a_1, a_2, \ldots a_p, b_1, b_2, \ldots, b_p)} e^{\sum_{j=1}^{p}[a_j \cos(j\theta) + b_j \sin(j\theta)]}, 0 \leq \theta < 2\pi \tag{12}$$

where

$$I_0 = I_0(a_1, a_2, \ldots a_p, b_1, b_2, \ldots, b_p) = \frac{1}{2\pi} \int_0^{2\pi} e^{\sum_{j=1}^{p}[a_j \cos(j\theta) + b_j \sin(j\theta)]} d\theta.$$

We note that for $p = 1$, the probability distribution (12) reduces to a unimode von Mises distribution and for $p_0 \leq p$ and $a_j = b_j = 0$ for $j \neq 0p$, the probability distribution (12) reduces to a p_0-mode von Mises distribution.

Assuming the distribution (12) for Θ, the torsional entropy of the system[1] is given by

$$S = k_B \left[\log 2\pi + \log I_0(a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_p) - \sum_{j=1}^{p} \frac{a_j}{I_0} C_j \right. \\ \left. - \sum_{j=1}^{p} \frac{b_j}{I_0} D_j \right] \tag{13}$$

where

$$\begin{aligned} C_j &= C_j(a_1, a_2, \ldots a_p, b_1, b_2, \ldots, b_p) \\ &= \frac{1}{2\pi} \int_0^{2\pi} \cos(j\theta) e^{\sum_{l=1}^{p}[a_l \cos(l\theta) + b_l \sin(l\theta)]} d\theta \end{aligned}$$

and

$$\begin{aligned} D_j &= D_j(a_1, a_2, \ldots a_p, b_1, b_2, \ldots, b_p) \\ &= \frac{1}{2\pi} \int_0^{2\pi} \sin(j\theta) e^{\sum_{l=1}^{p}[a_l \cos(l\theta) + b_l \sin(l\theta)]} d\theta. \end{aligned}$$

Given estimates of the parameters $a_1, a_2, \ldots, a_p$, $b_1, b_2, \ldots, b_p$, the estimates of the torsional entropy can be calculated.

---

[1] In the following the kinetic energy term is omitted for simplicity. Refer to Hnizdo *et al* [2002] for the full statistical-mechanical treatment of the problem.

Let $\theta_1, \theta_2, \ldots, \theta_n$ denote n random observations of the angle Θ. The log of the likelihood function is given by

$$\begin{aligned}\ln L = &- n \ln 2\pi - n \ln I_0(a_1, a_2, \ldots, a_p, b_1, b_2, \ldots, b_p) + \\ &\sum_{j=1}^{p} \sum_{i=1}^{n} \left[a_j \cos(j\theta_i) + b_j \sin(j\theta_i) \right]. \end{aligned} \tag{14}$$

Equating to 0 the partial derivatives of $\log L$ with respect to the $2p$ parameters, we obtain the following $2p$ likelihood equations which can be solved iteratively to obtain the maximum likelihood estimates of $a_1, a_2, \ldots a_p$, $b_1, b_2, \ldots, b_p$.

$$\frac{\partial \ln L}{\partial a_j} = -n \frac{C_j(a_1, a_2, \ldots a_p, b_1, b_2, \ldots, b_p)}{I_0(a_1, a_2, \ldots a_p, b_1, b_2, \ldots, b_p)} + \sum_{i=1}^{n} \cos(j\theta_i) = 0$$

$j = 1, 2 \ldots, p$.

$$\frac{\partial \ln L}{\partial b_j} = -n \frac{D_j(a_1, a_2, \ldots a_p, b_1, b_2, \ldots, b_p)}{I_0(a_1, a_2, \ldots a_p, b_1, b_2, \ldots, b_p)} + \sum_{i=1}^{n} \sin(j\theta_i) = 0$$

$j = 1, 2, \ldots, p$.

The value of p can be chosen by 'hit and miss' so that the fit is sufficient. Also, those parameters which have insignificant values can be dropped to remove an 'overfitting problem'. The authors used the software package MINUIT for obtaining the solution of the likelihood equations. The convergence of the iterative algorithm was rapid. As an illustration, they considered modeling the Ψ_1 angle of the AYPYD peptide in the folded conformation. ψ_1 is a ψ angle (on the backbone of the peptide N-C$_\alpha$-C-N, between C$_\alpha$ and C) of the first residue, in the case of AYPYD it is 'A' or the amino acid *alanine* [Demchuk *et al*, 1997a; Demchuk *et al*, 1997b]. The goodness of fit was verified by using chi-squared goodness of fit criterion. The fitted probability density function is given by

$$\begin{aligned} f(\psi_1) = &0.028 e^{0.715\cos(\psi_1) + 0.962\cos(3\psi_1) - 1.886\sin(\psi_1) - 1.623\sin(2\psi_1)}, \\ &0 \le \psi_1 < 2\pi. \end{aligned}$$

They also used this approach for modeling the distribution of the torsional angle of two halocarbon molecules. Conceptually, this approach has a natural extension to simultaneous modeling of two and more torsional angles, although finding the estimates of Fourier coefficients for modeling too many torsional angles is computationally expensive.

Singh *et al* [2002] proposed 2^{nd} and higher order nearest neighbor nonparametric estimates of entropy. These are based on the second and higher nearest neighbor distances between the sample points and provide competing estimators to the nearest neighbor estimate of entropy provided by Kozachenko and Leonenko [1987]. They established the asymptotic unbiasedness and consistency of the proposed estimators and investigated their performances for finite sample sizes for some standard distributions using Monte Carlo simulations. These estimators are used to evaluate the full internal entropy of methanol and diethyl ether.

4 References

Demchuk, E. and Singh, H. (2001). Statistical Thermodynamics of hindered rotation from computer simulations. *Molecular Physics*, **99**, 627-636.

Demchuk, E., Bashford, D. and Case, D.A. (1997a). Dynamics of a type VI reverse turn in a linear peptide in aqueous solution. *Folding & Design*, **2**, 35-46.

Demchuk, E., Bashford, D., Gippert, G.P. and Case, D.A. (1997b). Thermodynamics of a reverse turn motif. Solvent effects and side chain packing. *J. Mol. Biol.*, **270**, 305-317.

Hnizdo, V., Fedorowicz, A., Singh, H. and Demchuk, V. (2002). Statistical thermodynamics of internal rotation in a hindering potential of mean force obtained from computer simulations. In preparation.

Landau, L. D. and Lifshitz, E. M. (1980). *Statistical Physics*, Part 1, 3rd Edition, Oxford: Pergamon.

Karplus, M. and Kushick, J.N. (1981). Method for estimating the configurational entropy of macromolecules *Macromolecules*, Vol. 14, pp. 325.

Kozachenko, L. F. and Leonenko, N. N. (1987). Sample estimates of entropy of a random vector. *Problems of Information Transmission*, **23**, 95-101.

Levy, R.M., Karplus, M., Kushik, J. and Perahia, D. (1984). Evaluation of the configurational entropy for proteins: Application to molecular dynamics simulations of an α-helix, *Macromolecules*, Vol. 17, pp. 1370.

Singh, H., Hnizdo, V. and Demchuk, E. (2002). Probabilistic model for two dependent circular variables. Submitted for publication.

Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A. and Demchuk, E. (2002). Nearest neighbor estimates of entropy. Submitted for publication.

Singh, H., Hnizdo, V. and Demchuk, E. (2002). A Fourier series expansion of the potential function approach to probabilistic modeling of torsional angles. In preparation.

Devor, Linda

From: Musgrave, Katie
Sent: Monday, January 06, 2003 3:59 PM
To: Devor, Linda
Subject: Publication info requested

Linda: according to Gene Demchuk the "Stochastic Modeling of Rotational Degrees of Freedom in Molecules" is published as CD ROM (nonperiodical) Conference Proceedings of the Hawaii Int'l. Conference on Statistics, June 5-9, 2002, Honolulu, Hawaii. The editor was Kwok Ho. There are no page numbers due to CD ROM format.

Katie Musgrave

NIOSH/HELD/BEB

304-285-6256

◆◆◆

Stochastic modeling of rotational degrees of freedom in molecules. Proceedings of the Hawaii International Conference on Statistics, June 5-9, Honolulu, HI1-9.