

Theoretical and Computational Challenges in Entropy Evaluation of Macromolecules

H. Singh, J. Harner
Department of Statistics
West Virginia University
Morgantown, WV 26506

V. Hnizdo, E. Demchuk
Health Effects Laboratory Division
National Institute
for Occupational Safety and Health
Morgantown, WV 26505

Abstract

Evaluation of entropy is important in biological processes in order to predict the stability of a molecular conformation. The entropy evaluation requires probabilistic modeling of conformations in the internal coordinates. Since fluctuations in the rotational (torsional) angle coordinates make a pivotal contribution to the overall configurational entropy of the molecule, we review circular probability approaches for modeling the torsional angles. Since macromolecules such as proteins have a very large number of interdependent torsional angles and the distributions of many of them could be multimodal and even skewed, we discuss theoretically and computationally challenging problems that arise in the simultaneous modeling of these angles based on data from molecular dynamics simulations.

1 Introduction

Most complex biological processes are ultimately driven by elementary changes in the thermodynamic potential associated with the enthalpy and entropy of the system. Since life is only marginally stable, weak molecular interactions and the configurational freedom of the system (the entropy) often play a dominant role in a transition of one thermodynamic state to another. Many pivotal biological processes such as protein folding, intermolecular protein-protein interactions, protein-ligand interactions and others are known to be strongly dependent on variations in the configurational entropy. Therefore, a theoretical assessment of the entropy of a molecular system is an important biological problem that relates to system stability.

Probabilistic modeling of the torsional angles of the system is one of the cornerstones of the entropy evaluation procedure. It is a complex multidisciplinary problem that poses both theoretical and computational challenges. Karplus and Kushik (1981) and Levy et al. (1984) proposed modeling of torsional angles of large molecular systems (macromolecules) using a multivariate Gaussian distribution and then used the proposed model for entropy evaluation. Generally, the Gaussian distribution is inappropriate for any molecule with large fluctuations around rotatable bonds. Moreover, multiple peaks are commonly observed in the marginal distributions of torsional angles.

Demchuk and Singh (2001) introduced a new approach to the modeling of torsional angles which relies on circular statistics rather than linear statistics. Assume that the kinetic energy in a classical Hamiltonian is a conditional invariant of the

coordinates. Let $\Theta_1, \Theta_2, \dots, \Theta_m$ be the m internal curvilinear torsional angles of the system and let $f(\theta_1, \theta_2, \dots, \theta_m)$ be the joint configurational probability density function of $\Theta_1, \Theta_2, \dots, \Theta_m$. Then

$$f(\theta_1, \theta_2, \dots, \theta_m) = \frac{1}{Z(\beta)} e^{-\beta V(\theta_1, \theta_2, \dots, \theta_m)}, \quad (1)$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant, T is the temperature, $V(\theta_1, \theta_2, \dots, \theta_m)$ is the effective potential energy of the system when $\Theta_i = \theta_i$, $i = 1, 2, \dots, m$, and the configurational integral of the system, $Z(\beta)$, is chosen so that $f(\theta_1, \theta_2, \dots, \theta_m)$ is a probability density function, i.e.,

$$Z(\beta) = \int \int \dots \int e^{-\beta V(\theta_1, \theta_2, \dots, \theta_m)} d\theta_1 d\theta_2 \dots d\theta_m. \quad (2)$$

Let the Θ_i angles be independent and let each Θ_i have an l_i -mode von Mises distribution (Mardia and Jupp, 1999) with the probability density function

$$f_i(\theta_i) = \frac{1}{2\pi I_0(\kappa_i)} e^{\kappa_i \cos(l_i(\theta_i - \theta_{0i}))}, \quad -\pi < \theta_i \leq \pi, \quad (3)$$

where $\kappa_i > 0$, $i = 1, 2, \dots, m$ and $I_0(\cdot)$ is the modified Bessel function of order 0 defined by

$$I_0(\kappa) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{\kappa \cos(\theta)} d\theta. \quad (4)$$

Demchuk and Singh (2001) derived the following expression for the classical configurational torsional entropy of the system

$$S_c = k_B \left[m \ln(2\pi) + \sum_{i=1}^m \ln [I_0(\kappa_i)] - \sum_{i=1}^m \kappa_i \frac{I_1(\kappa_i)}{I_0(\kappa_i)} \right], \quad (5)$$

where $I_1(\cdot)$ is the modified Bessel function of order 1 and is given by

$$I_1(\kappa) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\theta) e^{\kappa \cos(\theta)} d\theta. \quad (6)$$

Thus given the values of the κ_i 's (or their estimates), the classical configurational torsional entropy (or its estimate) can be found using the tables of modified Bessel functions.

As a case study, they considered modeling the torsional angle of a methanol molecule by a 3-mode von Mises distribution and derived the following bathtub shaped probability density function of torsional energy $V = (V_0/2)(1 - \cos(3\phi))$ of the system, which is given by

$$g(v) = \frac{1}{\pi I_0(\kappa)} e^{\kappa \left(1 - \frac{2v}{V_0}\right)} v^{-1/2} (V_0 - v)^{-1/2}, \quad 0 \leq v \leq V_0. \quad (7)$$

Both the 3-mode von Mises and the bathtub shaped distributions, respectively, provided excellent fits to the torsional angle data and to the energy data obtained through molecular dynamics (MD) simulations.

MD is a powerful and currently used standard tool for investigation of structural and thermodynamic properties of molecular systems. Appreciation of computational methods as indispensable instruments for molecular analysis came from the fact that they have been elaborated to such an extent that they now often

produce more accurate results than experiments. In addition, MD provides much more information about the system of interest than any experiment does. Usually MD is started from an experimentally determined (X-ray or NMR) static three-dimensional structure of a molecule, or its quantum-mechanically calculated set of coordinates. Thus at first each atom is assigned a predetermined position in space. Then each atom is assigned a velocity that is randomly sampled from the Maxwellian distribution at a specific temperature and pressure. Atoms interact with each other according to the laws of physics, and therefore the whole system is able to evolve in time. Evolution is achieved by numerical propagation of Newtonian equations of motion at infinitesimal intervals, typically at each 10^{-15} sec. Characteristics of the evolving system are recorded. The time interval of recording is chosen so that during that period the system “forgets” about its previous state, i.e., the snapshots are independent of each other. A result of this procedure is a randomly sampled ensemble of the given molecular system simulated under specific well-controlled conditions; i.e., it is the full statistical-mechanical description of the process that is required by the Boltzmann-Gibbs theory.

Molecules generally have more than one torsional angle and often these angles are interdependent. Singh et al. (2001) proposed a probabilistic model on the torus for modeling two dependent circular variables which allows multiple peaks in the marginal distributions. In this model the joint probability density function of the two torsional angles Θ_1 and Θ_2 is given by

$$\begin{aligned}
 f(\theta_1, \theta_2) = & \\
 & C e^{\kappa_1 \cos[l_1(\theta_1 - \mu_1)] + \kappa_2 \cos[l_2(\theta_2 - \mu_2)] + \lambda \sin[l_1(\theta_1 - \mu_1)] \sin[l_2(\theta_2 - \mu_2)]}, \\
 & -\pi < \theta_1, \theta_2 \leq \pi,
 \end{aligned} \tag{8}$$

where $\kappa_1, \kappa_2 \geq 0$, $-\infty < \lambda < \infty$, $-\pi < \mu_1, \mu_2 \leq \pi$, l_1 and l_2 are positive integers, and C is a normalizing constant so that $f(\theta_1, \theta_2)$ is a probability density function. This distribution embeds a natural torus version of the bivariate normal distribution. They obtained expressions for the normalizing constant and the marginal variances and showed that the conditional distributions are von Mises distributions. The marginal distribution of $\Theta_1(\Theta_2)$ is either a $l_1(l_2)$ -mode von Mises-like or a $2l_1(2l_2)$ -mode symmetric distribution. They derive conditions on parameters which assure a specific shape. They applied this model for modeling two angles in methanol and in a short linear peptide which is a fragment of a protein.

2 Challenging Theoretical and Computational Problems

The distribution of torsional angles of many molecules do not have a symmetrical shape. The standard circular distributions discussed above are not suitable for modeling the distribution of such angles. Probabilistic modeling of these angles poses theoretical and computationally challenging problems. Hnizdo et al. (2001) proposed a Fourier series expansion of the potential function approach to modeling the probability distribution of such angles. First we discuss this approach for one such angle. We make a simple assumption that the energy, as a function of this angle, can be approximated by a finite number of terms in the Fourier series expansion. Let Θ denote such an angle and let $V(\theta)$ denote the torsional *potential* energy

of the molecule when the torsional angle takes the value θ . The Boltzman-Gibbs' probability density function of Θ is given by

$$f(\theta) = Ce^{-V(\theta)}, \quad (9)$$

where C is the normalizing constant. Consider the approximating finite Fourier series expansion of $V(\theta)$ given by

$$V(\theta) = \sum_{j=0}^p (a_j \cos(j\theta) + b_j \sin(j\theta)), \quad (10)$$

where a_j and b_j are fixed constants. Let $\theta_1, \theta_2, \dots, \theta_n$ denote the n random observations on angle Θ . The log of the likelihood function is given by

$$\ln(L) = n \ln(C) - \sum_{j=0}^p \sum_{i=1}^n (a_j \cos(j\theta_i) + b_j \sin(j\theta_i)). \quad (11)$$

They showed that $\ln(L)$ is a concave function in each of the parameters. In the evaluation of maximum likelihood estimates, the value of p can be chosen by 'hit and trial.' Also, parameters which have insignificant values can be dropped to remove the 'overfitting' problem. They used this approach to fit distributions to several torsional angles of a pentapeptide having 24 torsional angles using data from a molecular dynamics simulation of the pentapeptide (Demchuk et al., 1997). They used the software 'MINUIT' (James, 2001) and the convergence of the iterative algorithm was rapid. In the next section, we discuss an approach to developing fitting procedures for this problem using R (<http://www.r-project.org/>). Since macromolecules have a large number of angles, which are often dependent, simultaneous modeling of them using a Fourier series expansion of a multivariate potential function is in principle possible. However, the estimation of a very large number of parameters (Fourier coefficients) in the multivariate distribution is expected to be computationally challenging because of the large number of parameters in the case of macromolecules and also because of possible convergence problems of the iterative algorithms.

3 R as a Computing Environment

R provides a powerful environment for statistical computing, including the optimization of functions. We will use R, in essence as middleware, to compute maximum likelihood estimates of the parameters (e.g., the coefficients in the Fourier expansion) and to present the results to researchers.

3.1 Optimization Routines

MINUIT (James and Roos, 1975; James, 2001) is a system for function minimization developed at CERN—starting in 1967. It contains several minimization methods, which have changed and evolved over the years, to find (ultimately) a global minimum.

These minimization methods are logically connected so that the user can switch from one method to another to improve convergence. Currently, the minimization algorithms implemented are: a Monte Carlo search procedure (SEEK); a simplex method (SIMPLX) due to Nelder and Mead (1965); a variable metric method

(MIGRAD) due to Fletcher (1970). SEEK is used when a good starting point is unknown or when several local minima are expected, but it should be used primarily for exploration. MIGRAD, which requires first derivatives, is well behaved for roughly quadratic surfaces, but is not well behaved for complex surfaces. MIGRAD can automatically call SIMPLX if MIGRAD fails which can be fast even if the current estimates are far from the optimal values. Other global logic, such as bounding or fixing some of the parameters, is also available.

MINUIT is very powerful, but its control program is difficult and time consuming to use. We are searching for alternative optimization algorithms (e.g., the `ms` algorithm in R), which have been or could be converted to R packages or external routines. However, we like the flexibility and logic of MINUIT and currently we are exploring ways of porting it to R.

As a result, we are working on a hybrid approach which combines the features of R and MINUIT. The control programs in MINUIT will be replaced by R functions. The computational algorithms of MINUIT (written in Fortran) will be repositioned as a backend to R using foreign language interfaces. The optimization process will be controlled by R functions which will be successively modified to automate as much of the process as possible. As in MINUIT, covariance matrices of the parameter estimates will be computed whenever possible.

MINUIT currently is limited to 100 parameters and of these only 50 can be varied at a time. Strategies will be developed and implemented as R functions to automate the process of optimizing the likelihood function by fixing successive subsets of the parameters. This will allow moderately sized problems to be solved, but very large macromolecules will not be able to be accommodated.

The simulated data for each angle is often plotted in a histogram. The fitted density $\hat{f}(\theta)$ is then superimposed on the histogram; generally the fit has been good. R can easily accommodate this type of plot which lets us display the histogram and visualize the fitted density. In the multivariate case, \hat{f} will be a function of m torsional angles. It will be challenging to visualize the m -dimensional histogram of the angles and the superimposed fit. The fitted models will also be compared to various univariate and multivariate density estimates (e.g., Scott, 1992).

3.2 A Java/R Frontend

For the case of complex macromolecules, the probability density function could have many parameters and thus the computations can become very intensive. As a result, the R/MINUIT hybrid will be implemented on a large multi-processor system. Initially, X Windows will be used to access, control, and display the results. However, Java provides a better solution.

The R/Java interface prototyped by omegahat (<http://www.omegahat.org/>) allows R to be called from Java or vice versa. JavaStat (Xue and Harner, 2001) is being developed as a collaborative, distributed statistical system for doing basic statistical analyses and graphs. JavaStat was originally designed as an educational tool for teaching introductory statistics and thus it has limited analytical functionality. However, it is being extended by providing Java interfaces to R functions and thus its capabilities will be greatly extended. JavaStat has a spreadsheet interface to the data, which is formatted by an XML schema, and various graphical displays are available, including dynamic co-plots.

A mechanism will be developed to control R/MINUIT from JavaStat. This will allow researchers to model torsional angles remotely and collaboratively. Once the maximum likelihood estimates are passed from R to JavaStat, the data and fitted

density can be plotted. Various plotting systems will be explored, including Orca, linked to JavaStat (Lumley, 2001), and GGobi, linked to R (Lang, D.T. and Swayne, D.F., 2001).

4 Summary and Conclusions

The natural molecular coordinate system is intrinsically curvilinear due to kinematic constraints imposed on the atoms by covalent bonding. Therefore, an approach based on circular rather than linear statistics is more appropriate for modeling the joint probability distribution of coordinates, and thus more accurate thermodynamic calculations are possible. We have demonstrated an advantage of this new approach to modeling dihedral torsion angles of simple molecules. However, for macromolecules this approach implies challenging problems in the probabilistic modeling and in computations of parametric estimators of the models. We have discussed these problems in this paper along with several approaches to tackling these problems.

References

Demchuk, E. and Singh, H. (2001), “Statistical thermodynamics of hindered rotation from computer simulations,” *Molecular Physics*, Vol. 99, pp. 627–636.

Demchuk, E., Bashford, D., Gippert, G.P., and Case, D.A. (1997), “Dynamics of a type VI reverse turn in a linear peptide in aqueous solution,” *Folding and Design*, Vol. 2, pp. 35–46.

Fletcher R. (1970), “A new approach to variable metric algorithms,” *Comput. J.* Vol 13, pp. 317.

Hnizdo, V., Singh, H., Demchuk, E. (2001), “A Fourier series expansion of the potential function approach to probabilistic modeling of torsional angles,” preprint.

James, F. (2001), “MINUIT Minimization Manual,” CERN Program Library Long Writeup D506, Ver. 94.1.
(<http://wwwinfo.cern.ch/asdoc/WWW/minuit/minmain/minmain.html>)

James, F. and Roos, M. (1975), “MINUIT—A system for function minimization and analysis of the parameter errors and correlations,” *Comp. Physics Comm.* Vol. 10, pp. 343–367.

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

Lang, D.T. and Swayne, D.F. (2001), “GGobi meets R: an extensible environment for interactive dynamic data visualization,” *Proceedings of the 2nd International Workshop on Distributed Statistical Computing*, March 15-17, 2001, Technische

Universitt Wien, Vienna, Austria, (<http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/>)

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.

Lumley, T. (2001), "Orca [R [RJava]]," *Proceedings of the 2nd International Workshop on Distributed Statistical Computing*, March 15-17, 2001, Technische Universitt Wien, Vienna, Austria, (<http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/>)

Mardia, K. V. and Jupp, P. E. (1999), *Directional Statistics*, Wiley, New York.

Nelder, J.A. and Mead R. (1965), "A simplex method for function minimization," *Comput. J.* Vol. 7, pp. 308.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York.

Singh, H., Hnizdo, V. and Demchuk, E. (2001), "Probabilistic model for two dependent circular variables," communicated for publication.

Xue, H. and Harner, E. J. (2001), "JavaStat: A Distributed Statistical Computing Environment," *Proceedings of the 2nd International Workshop on Distributed Statistical Computing*, March 15-17, 2001, Technische Universitt Wien, Vienna, Austria. (<http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/>)