Using the Fisher-Bingham distribution in stochastic models for protein structure

John T. Kent*[1] and Thomas Hamelryck[2]

[1] University of Leeds
[2] University of Copenhagen

1 Introduction

Protein structure prediction is at present an unsolved problem, partly due to lack of a suitable energy function and partly due to the high number of conformational parameters involved. Many protein structure prediction methods therefore use simplified models (Buchete *et al.*, 2004). A popular approach is to represent a protein by its Cα atoms only, ie. by its Cα trace.

Thus a protein can be regarded as a sequence of edges in $\mathbb{R}^3$. To a first approximation the length of each edge is constant (in practice one unit of length is about 3.8 Å). The angles between successive edges determine the shape of the protein.

Here we introduce a probabilistic model of the geometry of a random protein's Cα trace. Such a model has many applications, like for example steering the proposal of protein conformations in a Markov Chain Monte Carlo (MCMC) simulation. The model is based on machine learning techniques, and is trained from a set of representative protein structures. Central to the model is the Fisher-Bingham distribution on the 3D sphere (Kent, 1982), an analogue of the bivariate normal equation in the plane.

2 Simulating the FB_5 distribution

The FB_5 (κ, β, R) distribution, where κ and β are real concentration parameters and R is a 3×3 orthogonal matrix representing orientation, was introduced in Kent (1982) and defines a statistical model on the unit sphere in $\mathbb{R}^3$. Its pdf in polar coordinates is given by

$$f(\theta, \phi) \propto \exp\{\kappa \cos\theta + \beta \sin^2\theta(\cos^2\phi - \sin^2\phi)\} \sin\theta, \tag{1}$$

where $\theta \in [0, \pi]$ denotes the colatitude and $\phi \in [0, 2\pi)$ denotes the longitude. Euclidean coordinates are defined by

$$u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = R \begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix}, \tag{2}$$

and we write $u \sim \mathrm{FB}_5(\kappa, \beta, R)$. The concentration parameters are usually required to satisfy

$$\kappa \geq 0, \quad 0 \leq \beta \leq \kappa/2, \tag{3}$$

and we shall restrict attention to this situation in this paper. In this setting the exponent $\{\kappa\cos\theta + \beta\sin^2\theta(\cos^2\phi - \sin^2\phi)\}$ is a nonincreasing function of $\theta \in [0, \pi]$ for each ϕ. (On the other hand, if $\beta > \kappa/2$, the pdf increases and then decreases in θ when $\phi = 0$.)

The FB_5 distribution was created to provide a spherical analogue for the bivariate normal distribution. The parameter β measures anisotropy. If $R = I$ in (1) the distribution is standardized so that the mode lies in the u_3 direction, and the principal axes are given by the u_1 and u_2

axes, respectively. Under large concentration, the distribution follows an asymptotic bivariate normal distribution when orthogonally projected onto the tangent plane of the sphere.

For the purposes of simulation it is helpful to use an equal area projection. Set

$$x_1 = r\cos\phi,\ x_2 = r\sin\phi,\ \text{where } r = \sin(\theta/2), \tag{4}$$

so that $(2x_1, 2x_2)$ represents an equal-area projection of the sphere.

In (x_1, x_2) coordinates, the Jacobian factor $\sin\theta$ disappears and the pdf (with respect to $dx_1\,dx_2$ in the unit disk $x_1^2 + x_2^2 < 1$) takes the form

$$\begin{aligned} f(x_1, x_2) &\propto \exp\{-2\kappa r^2 + 4\beta(r^2 - r^4)(\cos^2\phi - \sin^2\phi)\} \\ &= \exp\{-2\kappa(x_1^2 + x_2^2) + 4\beta[1 - (x_1^2 + x_2^2)(x_1^2 - x_2^2)]\} \\ &= \exp\{-\frac{1}{2}[ax_1^2 + bx_2^2 + \gamma(x_1^4 - x_2^4)]\}. \end{aligned} \tag{5}$$

where the new parameters

$$a = (4\kappa - 8\beta), \quad b = (4\kappa + 8\beta), \quad \gamma = 8\beta \tag{6}$$

satisfy $0 \le a \le b$ and $\gamma \le b/2$. Here we have used the double angle formulas, $\cos\theta = 1 - 2\sin^2(\theta/2),\ \sin\theta = 2\sin(\theta/2)\cos(\theta/2)$.

Note that the pdf splits into a product of a function of x_1 alone and x_2 alone. Hence x_1 and x_2 would be independent except for the constraint $x_1^2 + x_2^2 < 1$. Our method of simulation, as sketched below, will be to simulate $|x_1|$ and $|x_2|$ separately by acceptance-rejection using a (truncated) exponential envelope, and then additionally to reject any values lying outside the unit disk.

Wood (1987) has also developed a simulation algorithm for the Fisher-Bingham distribution. His method is more general because it includes a wider range of parameter values and also includes the more general FB_6 distribution. However, the method described here is simpler to implement when (3) is satisfied.

The starting point for our simulation method is the simple inequality

$$\frac{1}{2}(\sigma|w| - \tau)^2 \ge 0 \tag{7}$$

for any parameters $\sigma,\ \tau > 0$ and for all w. Hence

$$-\frac{1}{2}\sigma^2 w^2 \le \frac{1}{2}\tau^2 - \sigma\tau|w|. \tag{8}$$

After exponentiation, this inequality provides the basis for simulating a Gaussian random variable from a double exponential random variable by acceptance-rejection. But for our purposes some refinement is needed.

For x_1 we need to apply (8) twice, first with $\sigma = \gamma^{1/2},\ \tau = 1$ and $w = x_1^2$, and second with $\sigma = (a + 2\gamma^{1/2})^{1/2},\ \tau = 1$ and $w = x_1$, to get

$$\begin{aligned} -\frac{1}{2}(ax_1^2 + \gamma x_1^4) &\le \frac{1}{2} - \frac{1}{2}(a + 2\gamma^{1/2})x_1^2 \\ &\le c_1 - \lambda_1|x_1| \end{aligned} \tag{9}$$

where

$$c_1 = 1, \quad \lambda_1 = (a + 2\gamma^{1/2})^{1/2}. \tag{10}$$

To develop a suitable envelope for x_2 recall that $0 \leq 2\gamma \leq b$. To begin with suppose $b > 0$. From (8) with $\sigma = (b-\gamma)^{1/2}$, $\tau = (b/(b-\gamma))^{1/2}$, and $w = x_2^2$,

$$-\frac{1}{2}(bx_2^2 - \gamma x_2^4) \leq -\frac{1}{2}(a-\gamma)x_2^2 \leq c_2 - \lambda_2|x_2| \tag{11}$$

where

$$c_2 = b/\{2(b-\gamma)\} \leq 1, \quad \lambda_2 = b^{1/2}. \tag{12}$$

If $b = 0$ (and so $\gamma = 0$) then (11) continues to hold with $\lambda_2 = 0$ and $c_2 = 0$.

3 A random walk protein model

Consider a sequence of vertices $\{v_i,\ i = 1, \ldots, n\}$ in $\mathbb{R}^3$ representing the Cα trace of a protein. Edges can be defined by $e_i = v_i - v_{i-1}$. By assumption the e_i all have unit size.

Two successive edges e_{i-1} and e_i determine a 3×3 orthogonal matrix $G = [g_1, g_2, g_3]$ defining a frame of reference at vertex i as follows:

$$g_3 = e_i, \quad g_1 = e_{i-1} - (e_{i-1}^T e_i)\, e_i, \quad g_2 = g_3 \times g_1.$$

Our first model for the protein structure will be a third-order Markov process. Consider an FB_5 distribution $FB_5(\kappa, \beta, R)$ with fixed parameters. Then $v_{i+1} = v_i + e_{i+1}$ is simulated by

$$G^T e_{i+1} \sim FB_5(\kappa, \beta, R).$$

where the different FB_5 simulations are independent for each i. The process is third-order Markov because of the need to determine a frame of reference at each vertex.

4 An HMM protein model

The previous model is a bit too simplistic to be useful in practice. Hence we let the parameters (κ, β, R) vary according to a hidden Markov model (HMM) with a finite number of states. We call this HMM with discrete hidden nodes and observed FB5 nodes FB5-HMM. The discrete hidden nodes of the FB5-HMM can be considered to model a sequence of fine-grained 'secondary structures'. The observed FB5 nodes translate these 'secondary structures' into corresponding angular distributions.

The FB5-HMM was implemented using Mocapy (Hamelryck, 2004). Mocapy is a parallelized Dynamic Bayesian Network (Jordon, 1998) toolkit that supports Discrete, (Multidimensional) Gaussian, Dirichlet, Von Mises-Fisher and FB5 nodes. Mocapy optimizes a model's parameters by Stochastic Expectation Maximization using Gibbs sampling to infer the values of the hidden nodes (Gilks *et al.*, 1996).

For training FB5-HMM, we used a set of 1600 representative protein structures. The size of the discrete hidden nodes in the HMM was 50. The calculations were done on the 240 CPU cluster of the Bioinformatics center. Preliminary results indicate that the trained FB5-HMM is an excellent model of a protein's Cα trace.

5 Possible applications

In general, the FB5-HMM model can be used to generate 'random proteins', ie. sequences of Cα atoms whose consecutive orientations resemble those of real proteins. This is where simulation

of the FB5 distribution as presented in this article comes into play. Of course, the model does not automatically lead to compact structures or structures that avoid self-intersections. However, it is relatively easy to generate structures that broadly resemble real proteins by introducing some rejection criteria for the sampled angles.

In particular, an obvious application of the method described here is in MCMC simulations of simplified protein models. The model can be used to generate protein conformations that can be accepted or rejected according to some energy function. Another application is in loop modelling: the model can be used to steer the generation of loops, for example in homology modelling. In this case, it is ensured that the proposed loops will adopt reasonably realistic conformations. The model can also be used to generate decoy structures that adopt realistic backbone conformations: decoys are widely used to evaluate energy function that are used for protein structure prediction.

6 Acknowledgements

TH is supported by a Marie Curie Intra-European Fellowship within the 6th European Community Framework Programme. Mocapy and the FB5-HMM model were developed in the laboratory of Prof. Anders Krogh, Bioinformatics Center, Institute of Molecular Biology and Physiology, University of Copenhagen, whose support is acknowledged.

References

Buchete, N. V., Straub, J, E, and Thirumalai, D. (2004). Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.* **14**, 225–232.

Gilks, W. R., Richardson, S. T. and Spiegelhalter, D. J. (1996). *Markov Chain Monte-Carlo in Practice*. Chapman and Hall, London.

Hamelryck, T. (2004). Mocapy: A parallelized toolkit for learning and inference in Dynamic Bayesian Networks. `https://sourceforge.net/projects/mocapy/`

Jordan, M. I. (1998). *Learning in Graphical Models*. MIT Press.

Kent, J. T. (1982). The Fisher-Bingham distribution on the sphere. *J Royal Statist. Soc.* **B 44**, 71–80.

Wood, A. T. A. (1987). The simulation of spherical distributions in the Fisher-Bingham family. *Commun. Statist. Simulation and Computation* **16**, 885–898.