

# Statistical Characterization of Protein Ensembles<sup>1</sup>

Diego Rother,<sup>2</sup> Guillermo Sapiro,<sup>2</sup> and Vijay Pande<sup>3</sup>

**Keywords:** protein ensembles, maximum likelihood, minimum entropy, kernel density estimation, bootstrap.

## 1 Introduction

A single structure is often used to represent a protein. This can be considered natural when the structure is assumed to be rigid, but protein structure is known to fluctuate in physiological conditions. This fact, overlooked for simplicity in many situations, may be advantageously accounted for at times when high resolution techniques for structure determination are available. Even if the “true” structure *were* fixed and unique, the uncertainty in its determination by the measurement of some property (i.e. diffraction, magnetic resonance, etc.) also produces variability, since those methods generally optimize a model to fit the observations, a process prone to find multiple local minima. A similar situation arises when simulations are used. The modeled energy landscape is populated with multiple local minima close to the global minimum. As a result of this and other intrinsic simulation characteristics (e.g. randomness), multiple representatives are possible. This fact was postulated to explain the robustness of the native state and suggested as a way to predict it for certain protein classes [3].

In applications where the fluctuations can not be ignored, and moreover, are to be favorably exploited, how should they be represented and incorporated into the calculations? One way is to represent the protein structure not as a single conformation but as a finite set of conformations, corresponding to the different “observations” of its state. We propose a different approach, consisting of estimating a probability density function (pdf) from the available samples of the state, and using this pdf to represent the ensemble.

This representation is also advantageous for at least two key reasons: 1) it may allow to see things that were hidden when the ensemble was regarded as a set of discrete conformations (e.g., the modes); and 2) provides a natural framework to perform certain operations (e.g. compare ensembles of the same protein that were obtained by different methods, or determine the probability that a particular conformation belongs to the ensemble [6]). An important by-product of the approach is that it provides an idea of the completeness of the sample in representing the space of conformations that the protein adopts.

In the next section we give a brief description of the mathematical and computational method developed to accomplish this task and in the last section we describe some of the results obtained.

## 2 Methods.

For simplicity we restrict our attention in this presentation only to the backbone of the protein, described by the usual *phi* ( $\phi$ ) and *psi* ( $\psi$ ) angles. Consequently, each conformation is represented by a vector of angles ( $\vec{x} = [\phi_1 \dots \phi_d \psi_1 \dots \psi_d]$ ) or to simplify the notation  $\vec{x} = [x_1 \dots x_d]$  in a  $d$  dimensional space ( $d=2(r-1)$  where  $r$  is the number of residues in the chain). Our goal is to estimate the probability density function  $p(\vec{x})$ . To do that, we first note that it can be written as

$$p(x_1, x_2, \dots, x_d) = p(x_{i_1} | x_{i_2}, \dots, x_{i_d}) \cdot p(x_{i_2} | x_{i_3}, \dots, x_{i_d}) \dots p(x_{i_d}) \quad (1)$$

which splits the joint density  $p(\vec{x})$  into  $d$  conditional densities.

To estimate these high dimensional conditional densities, and circumvent the “curse of dimensionality,” we consider that one angle of the conformation is only affected by the previous and following angles along the chain, and perhaps by some angles in its spatial proximity. Therefore, we can assume that each coordinate is only related to a small set of other coordinates ( $n$ ), and more important, is “independent” of the rest, and thus only a small subset of the variables should be included in the conditioning side of the conditional densities in (1), resulting in

---

<sup>1</sup> Work supported by ONR, DARPA and NSF.

<sup>2</sup> Department of Electrical and Computer Engineering, University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455, USA. E-mail: [diroth,guille@ece.umn.edu](mailto:diroth,guille@ece.umn.edu)

<sup>3</sup> Chemistry Department, Stanford University, Stanford, CA 94305-5080, USA.

$$p(x_1, x_2, \dots, x_d) \approx p\left(x_{i_1} \mid \underbrace{x_{i_1^1}, \dots, x_{i_1^n}}_{C_{i_1}}\right) \cdot p\left(x_{i_2} \mid \underbrace{x_{i_2^1}, \dots, x_{i_2^n}}_{C_{i_2}}\right) \dots p(x_{i_d}) \quad (2)$$

This reduces the problem to first finding which are those densities (factors on the right of Equation (2)) to be estimated, and then estimating them. Note that we do not assume knowledge of which are the  $n$  other angles that affect each angle, this is also computed as part of the algorithm. Since the number of maximum dependencies for a variable ( $n$ ) is not known a priori, a third problem that arises is to define and apply a criterion to define  $n$  in a sensible way.

These three problems are tackled using tools from the areas of statistics and information theory. We employed kernel smoothing (KS) [4] and maximum likelihood or minimal entropy cross-validation (CV) [5] to estimate the densities; maximum likelihood again (ML) to select the factors in Equation (2), and bootstrap (B) [1] to choose the dimension ( $n$ ) of the densities. Genetic algorithms (GA) are used to do some of the optimization. A coarse diagram of the process is given in Figure 1. Further details are presented in our extended report.

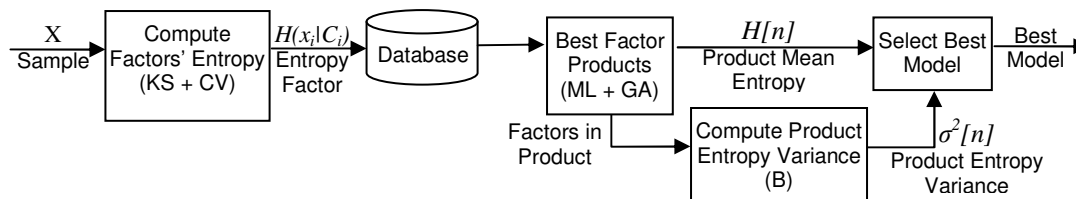


Figure 1. A reduced block diagram of the method proposed.

### 3 Results.

We first constructed a toy example to test the principle. It was designed to have the variables depending on at most three other variables. The score in the following graphs is a measure of how well the model represents the data, lower scores being better. This score is based on entropy computations obtained from the statistical tools mentioned above. Figure 2a shows that as expected, including more than three variables does not improve the fitting. Figure 2b shows the method applied to an ensemble consisting of 481 conformations of the  $\beta$ -hairpin tryptophan zipper [2], a peptide having 12 residues (22 dimensional pdf). It can be seen in this figure that for the current sample the benefit obtained from the new dependencies accounted for by including more than two conditioning variables is smaller than the harm done by the additional smoothing required. This suggests that the number of sample points in the ensemble does not need to grow exponentially with the number of residues in the peptide, but only with the number of those actually affecting each other. Applications are included in the extended report. These include finding the probability that a given conformation belongs to the ensemble of simulated folded proteins (using data from [2]). We also show the particular torsion angles that are strongly correlated, an immediate consequence of our approach.

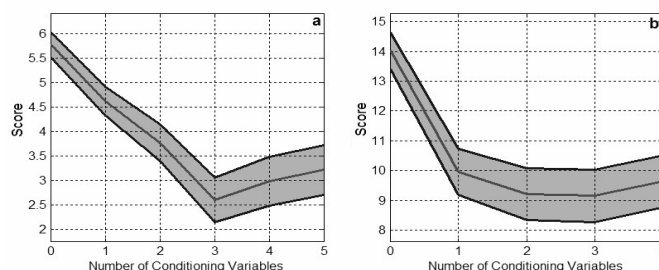


Figure 2. The method applied to two different datasets: a) a toy example and b) a very small protein having 12 residues. In both cases the band represents the mean score with a band of three standard deviations on each side.

### References

- [1] Efron, B. and Tibshirani, R.J. 1993. *An introduction to the bootstrap*. Chapman and Hall/CRC.
- [2] Pande, V.S. and Stanford University. 2000-2004. Folding@home distributed computing. <http://folding.stanford.edu/>
- [3] Shortle, D., Simons, K.T. and Baker, D. 1998. Clustering of low-energy conformations near the native structures of small proteins. *Proceedings of the National Academy of Sciences, USA*. Vol. 95, 11158-11162.
- [4] Silverman, B.W. 1986. *Density estimation for statistics and data analysis*. Chapman and Hall/CRC.
- [5] Viola, P.A. 1995. Alignment by maximization of mutual information. *Doctoral dissertation*. Massachusetts Institute of Technology.
- [6] Zagrovic, B., Snow, C.D., Khalid, S., Shirts, M.R. and Pande, V.S. 2002. Native-like mean structure in the unfolded ensemble of small proteins. *Journal of Molecular Biology*. Vol. 323, 153-164.