Guiding Particle Filtering with Marginal Approximations: an Application in Protein Image Interpretation

Frank DiMaio[1], Ameet Soni[1], Jude Shavlik[1], George Phillips[2,1]
[1]UW-Madison Computer Sciences Department,
[2]UW-Madison Biochemistry Department
Madison, WI 53706
[dimaio,shavlik,soni]@cs.wisc.edu, phillips@biochem.wisc.edu

We investigate the use of particle filtering [1] to automatically identify protein structures in electron density maps. A protein – a linear chain of amino acids that folds into some specific 3D conformation – consists of a repeating four-atom *backbone* motif with *sidechains* hanging off at uniform intervals; each of the twenty naturally occurring amino acids has a different sidechain. The electron density map, analogous to a 3-dimensional picture of a protein, is produced as the final result of x-ray crystallography. Interpreting these maps, that is, locating all the protein's atoms in these complex 3D images is often time consuming, requiring a crystallographer spend weeks to months tediously placing each atom.

Our previous work [2, 3] employs probabilistic inference to compute the marginal distribution of each amino-acid's 3D location on a grid. However, several simplifications are made by this model. First, our previous model identifies the location of just a single atom in each amino-acid, the alpha carbon (C_α). Biologists are interested in not just the position of each C_α, but in the location of each of the four to fourteen non-hydrogen atoms in each amino acid. Second, our model places these C_α's on a grid, typically with 1Å grid spacing. However, interatomic distances are known to much greater accuracy: the C_α–C_α bond is always 3.8Å, with a standard deviation of less than 0.1Å. By forcing C_α's to lie on grid point, we get a predicted protein structure that may not be physically feasible.

To address these shortcomings, and produce the most likely physically feasible all-atom protein model (or set of models), we have investigated the use of particle filtering (PF). Statistical importance resampling (SIR) [1, 4] – a particle filtering method – approximates some posterior probability distribution over a state sequence $x_{0:N}$ given observations $y_{0:N}$ as the sum of a finite number of point estimates $x^i_{0:N}$, each with weight w_i such that $\sum_{i=0}^{N} w_i = 1$. Assuming $x_{0:N}$ is a Markov process, we get the recursion:

$$p(x_{0:k}|y_{0:k}) \propto p(y_k|x_k) \cdot p(x_k|x_{k-1}) \cdot p(x_{0:(k-1)}|y_{0:(k-1)}) \tag{1}$$

SIR is based on the assumption that $p(y_k|x_k) \cdot p(x_k|x_{k-1})$ above is hard to sample directly, but easy to evaluate up to proportionality. If there is another distribution $q(x_k|x_{k-1}, y_k)$ from which we can directly sample, then we can use q to generate each x^i_k from x^i_{k-1} and y^i_k, then reweight each particle as the ratio of $p(x)$ to $q(x)$,

$$wt^i_k \propto wt^i_{k-1} \cdot \frac{p(y^i_k|x^i_k) \cdot p(x^i_k|x^i_{k-1})}{q(x^i_k|x^i_{k-1}, y^i_k)} \tag{2}$$

For density-map interpretation, each x_k is the position of every atom in amino acid k. We want to find the conformation $x_{0:N}$ that best explains the observed map. To simplify somewhat, we parameterize x_k as a C_α translation b_k, a rotation r_k and a sidechain orientation s_k. We consider only a finite number of distinct sidechain conformations; thus, s_k is simply an index into a database of known sidechain 3D structures. The probability of adding a residue in a specific conformation, $p(y_k|x_k) \cdot p(x_k|x_{k-1})$ is easily computed for some placement of atoms x_k, but difficult to sample: the first term is based on a measure used by crystallographers, the *R-factor* – which measures how well a map is explained by a model – while the second term is derived from previously solved structures.

Topic: prediction and sequence modeling
Preference: oral/poster

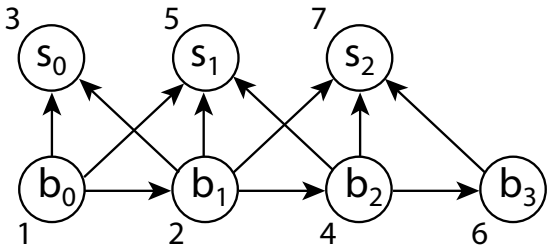

Figure 1: Conditional dependencies in sidechain (s_k) and C_α (b_k) layout. Numbers indicate the order in which labels are sampled. Rotations (r_k, not shown) are uniquely determined given s_k and $\langle b_{k-1}, b_k, b_{k+1}\rangle$.

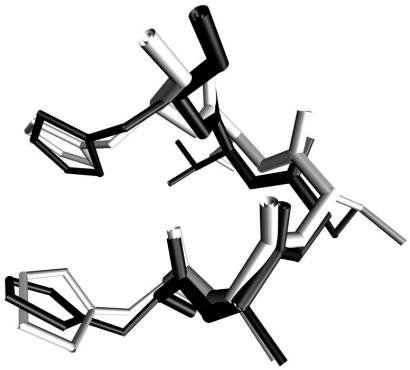

Figure 2: A comparison of the true structure (white) of a 5-amino-acid protein fragment and the highest-weight particle from a 500-particle run (black). The backbone is indicated by the thicker segments.

Our importance function $q(x_k|x_{k-1}, y_k)$ separates sampling the C_α translation of each amino acid and the sidechain conformation. Figure 1 illustrates how the protein structure is grown. At each step, we first sample C_α position b_{k+1} given b_k as the product of b_{k+1}'s probability based on observed conformations of $\langle b_k, b_{k+1}\rangle$, and the approximate marginal $\hat{p}_{k+1}(b_{k+1})$,

$$q(b_{k+1}|b_k) = p_{kinematics}(b_{k+1}|b_k) \cdot \hat{p}_{k+1}(b_{k+1}) \tag{3}$$

Once we have chosen location b_{k+1}, we pick a sidechain conformation s_k and orientation r_k. The probability of some sidechain conformation is computed from the correlation coefficient between the sidechain and the map: for each sidechain and orientation, we compute the probability p_k that the correlation coefficient was generated by chance; each sidechain conformation's probability is proportional to $(1 - p_k)/p_k$.

Next, given the triple $\langle b_{k-1}, b_k, b_{k+1}\rangle$ and the evidence (the density map), we sample r_k and s_k. Sampling from this distribution is straightforward; there are a finite number (usually 100 or so) of s_k's which we can completely enumerate, and – given s_k and $\langle b_{k-1}, b_k, b_{k+1}\rangle$ – there is only a single r_k with non-zero probability. Once amino acid k is sampled, the weights of each particle are updated as in Equation 2.

Using the marginals to guide sampling requires significantly fewer particles to recover an accurate structure than either using the priors to guide our search, or using kinematics alone. Figure 2 compares the true structure and the highest-weight predicted particle on a short protein segment. The error over this segment is very close to the average error over the entire protein.

Preliminary results using this method are very promising: in one protein backbone error is reduced from 1.1Å RMSd to 0.8Å RMSd; in another backbone error is reduced from 2.1Å RMSd to 1.8Å RMSd while the portion of the protein identified increases from 91% to 100%. Perhaps more importantly, these traces return a physically feasible all-atom model.

References

[1] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp (2001). A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. *IEEE Trans. of Signal Processing.*

[2] F. DiMaio, J. Shavlik and G. Phillips (2006). A probabilistic approach to protein backbone tracing in electron density maps. *Proc. ISMB.*

[3] F. DiMaio, J. Shavlik and G. Phillips (2006). Tracing protein backbones in electron density maps using a Markov random field model. *Snowbird Learning Workshop.*

[4] A. Doucet, S. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comp.*