# Graphical models and directional statistics capture protein structure

Wouter Boomsma[1], John T. Kent[2],
Kanti V. Mardia[2], Charles C. Taylor[2] & [1] Thomas Hamelryck*[1]

[1] Bioinformatics Centre, University of Copenhagen
[2] Department of Statistics, University of Leeds

## 1 Introduction

One of the major unsolved problems in modern day molecular biology is the protein folding problem: given an amino acid sequence, predict the overall three-dimensional structure of the corresponding protein. It has been known since the seminal work of Anfinsen (1973) in the early seventies that the sequence of a protein encodes its structure, but the exact details of the encoding still remain elusive. Since the protein folding problem is of enormous practical, theoretical and medical importance, and in addition forms a fascinating intellectual challenge, it is often called the holy grail of bioinformatics.

The conformational space potentially accessible to a protein is vast, and a brute force enumeration of all possible conformations to pinpoint the minimum energy conformation is computationally (and in fact also physically) impossible (Levinthal, 1969). Therefore, exploring the conformational space of a protein is typically done using a divide-and-conquer approach: plausible protein conformations are generated using a conformational sampling method, and the sampled conformations are accepted or rejected using some kind of energy function. In addition, protein structure prediction methods often make use of simplified models, where each amino acid in the polypeptide chain is represented by one or a few points in space.

However, efficient sampling of plausible protein structures that are compatible with a given amino acid sequence is a long standing open problem. We provide a solution to this problem by constructing probabilistic models of protein structure that represent the joint probability distribution of sequence and local protein geometry. The construction of these models becomes feasible by combining graphical models like Hidden Markov Models (HMMs) or Dynamic Bayesian Networks (Ghahramani, 1997) with directional statistics (Mardia and Jupp, 2000), which leads to tractable models that nonetheless represent protein structure in continuous space. We developed two models: the first dealing with C$\alpha$ geometry, the other one with full backbone geometry.

## 2 FB5-HMM: A model of C$\alpha$ geometry

A protein is a linear chain of amino acids. By representing each amino acid as a single point, corresponding to the C$\alpha$ atom, one obtains the so-called C$\alpha$ trace of the protein. The distance between two consecutive C$\alpha$ atoms can be considered fixed (about 3.8 Å), and as a result, the geometry of the C$\alpha$ trace can be parameterised by a sequence of $N-3$ dihedral angles (called $\tau$) and $N-2$ angles (called $\theta$) (Levitt, 1976; Oldfield and Hubbard, 1994).

By interpreting the $(\theta, \tau)$ angles as polar coordinates, a C$\alpha$ trace can also be fully described by a sequence of three-dimensional unit vectors. Therefore, a convenient way to construct a probabilistic model of the C$\alpha$ trace of proteins is to use an HMM that outputs amino acid symbols, secondary structure symbols ($\alpha$-helix, $\beta$-strand and coil), and unit vectors. Probability distributions over unit vectors are in the realm of directional statistics, which is concerned with the statistics of angles, orientation and directions (Mardia and Jupp, 2000). We used the 5-parameter Fisher-Bingham distribution on the unit sphere to represent the unit vectors (Kent, 1982; Kent and Hamelryck, 2005).

The resulting HMM, called FB5-HMM is shown in Fig. 1. The joint probability distribution of amino acid sequence $A$, secondary structure sequence $S$ and angle sequence $X$ is given by:

$$P(A, S, X) = \sum_H P(A \mid H)P(S \mid H)P(X \mid H)P(H)$$

where the sum runs over all possible hidden node sequences $H$. The parameters of the FB5 distribution are conditioned on the value of the hidden node.

Using an HMM with multiple outputs allows challenging operations like sampling a sequence of backbone angles given an amino acid and secondary structure sequence to be computed in an efficient way (Cawley and Pachter, 2003). The use of directional statistics makes it possible to represent protein structure in continuous space, without the need of the usual discretisations.

FB5-HMM was implemented in our Dynamic Bayesian Network toolkit Mocapy (Hamelryck, 2004) using using a dataset of more than 1400 protein structures. Analysis of FB5-HMM shows that the model captures protein structure extremely well, this for example in terms of angle content and secondary structure length distributions. Fig. 2 shows five typical C$\alpha$ trace samples for the sequence $\mathtt{V}_{15}\mathtt{INGKV}_{15}$. Valine (V) has a strong preference for the $\beta$-strand conformation, while the central INGK sequence is a sequence that is typical for a $\beta$-turn. These conformational preferences are clearly reflected in the sampled structures.

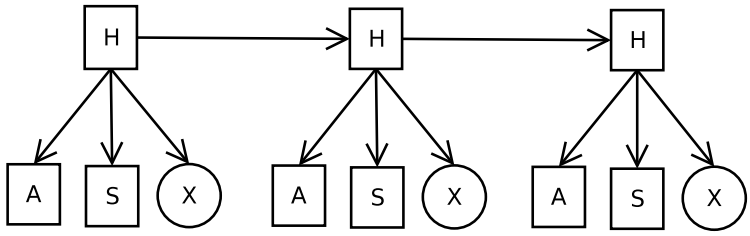



Figure 1: Conditional dependency diagram of the two HMMs described in the article. The HMM shown here corresponds to a sequence of length three. Arrows represent conditional dependencies. Each node represents a random variable: $H$: discrete hidden node; $A$: discrete amino acid node; $X$: continuous angle node. The latter node is an FB5 node for the C$\alpha$ model, and a Torus node for the full backbone model.

# 3 Torus-HMM: A model of full backbone geometry

The C$\alpha$ trace is a convenient way to represent proteins. It is however a simplified representation of the protein backbone: in an actual protein, consecutive C$\alpha$ atoms are not connected by physical bonds, but are separated by a nitrogen and a carbon atom (the backbone is a sequence of N, C$\alpha$ and C' atom triplets). While the C$\alpha$ representation successfully captures the topology

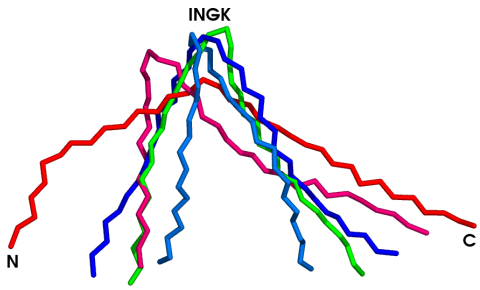
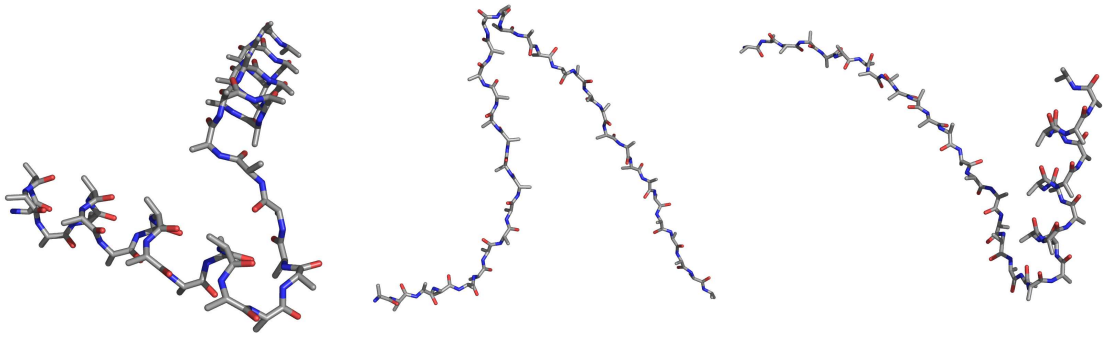



Figure 2: Five superimposed C$\alpha$ traces sampled for the sequence $\mathtt{V}_{15}\mathtt{INGKV}_{15}$. The N- and C-terminal ends, and the approximate position of the `INGK` sequence stretch are indicated. Figures 2 and 3 were made using PyMol (`http://pymol.sourceforge.net/`).

of the overall backbone, it contains no information on the position of these additional atoms, which is often necessary for the calculation of detailed energy functions.

Since both the bond angles, the dihedral angles associated with the peptide bonds, and the physical bond lengths between the various atoms are approximately fixed, also this model has two angular degrees of freedom for each amino acid. These are normally referred to as the $\phi$ and $\psi$ angles. Unlike the C$\alpha$ model, these angles are *both* dihedral angles, and thus both range from $-\pi$ to $\pi$. Pairs of such angles correspond to points on the surface of a unit torus and given the approach described for the FB5-HMM, the only additional requirement to model a full-atom backbone is a family of Gaussian-like distributions on the surface of a torus. For this purpose, we use the cosine model, which was recently proposed by Mardia, Subramaniam and Taylor (Mardia *et al.*, 2006). The optimal parameters of the resulting torus-HMM were found using the method described in the previous section.

Since the models described in this paper include information on secondary structure, they should be able to capture their corresponding angular preferences. Figure 3 shows the structures generated by the torus-HMM if we sample ($\phi$, $\psi$) angle sequences given a fixed secondary structure. We clearly observe that the secondary structure elements we use as input have significant influence on the produced local structure. Since the field of secondary structure prediction is well-developped, the ability to add secondary structure as input to our model makes it possible to use predicted secondary structure as a guideline for what type of angular preferences you expect to see in different parts of your unknown protein structure - and thereby reduce the size of the search space.



(a) (b) (c)

Figure 3: Structures sampled using the torus-HMM from a given secondary structure. (a) 15 $\alpha$-helix - 5 coil - 15 $\alpha$-helix, (b) 15 $\beta$-sheet - 5 coil - 15 $\beta$-sheet, (c) 15 $\beta$-sheet - 5 coil - 15 $\alpha$-helix. The N-termini are on the left.

# 4 Conclusions

The marriage of graphical models and directional statistics proved extremely fruitful for the development of probabilistic models of the local structure of proteins. By using an HMM with directional output (points on the unit sphere or on the torus) it becomes possible to construct a joint probability distribution over protein sequence and structure. Importantly, protein structure can be represented in a geometrically natural, continuous space. The two HMMs presented here will be used for the proposal of plausible protein geometries in a protein structure prediction method. The future of methods based on combining graphical models and directional statistics in bioinformatics looks bright, as these methods are powerful, computationally tractable and conceptually elegant.

## Acknowledgements

TH is supported by a Marie Curie Intra-European Fellowship in the 6th European Community Framework Programme. WB was supported by the Lundbeckfond (`www.lundbeckfonden.dk`).

## References

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223-230.

Cawley, S.L. and Pachter, L. (2003). HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19(2)**, II36-II41.

Ghahramani, Z. (1997). Learning Dynamic Bayesian Networks. *Lecture Notes in Computer Science*, **1387**, 168-197.

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

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

Kent, J. and Hamelryck, T. (2005). Using the Fisher-Bingham distribution in stochastic models for protein structure. In: S. Barber, P.D. Baxter, K.V. Mardia and R.E.Walls (*eds.*), *Quantitative Biology, Shape Analysis, and Wavelets*. University of Leeds Press, Leeds. 57-60.

Levinthal, C. (1969). How to Fold Graciously, In: J.T.P. DeBrunner and E. Munck (*eds.*), *Mössbauer spectroscopy in biological systems*. U. of Illinois Press, Illinois. 22-24.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, **104**, 59-107.

Mardia, K.V. and Jupp, P. (2000). Directional Statistics. Wiley, Chichester, 2nd ed.

Mardia, K.V., Taylor, C.C. and Subramaniam, G.K. (2006). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. Submitted.

Oldfield, T.J. and Hubbard, R.E. (1994). Analysis of C$\alpha$ geometry in protein structures. *Proteins*, **18**, 324-37.