On the reference ratio method and its application to statistical protein structure prediction

Thomas Hamelryck[1], John Haslett[2], Kanti Mardia[3], John T. Kent[3],
Jan Valentin[1], Jes Frellsen[1], Jesper Ferkinghoff-Borg[4]

[1]The Bioinformatics center, University of Copenhagen, Denmark
[2]School of Computer Science and Statistics,Trinity College, Dublin
[3]Department of Statistics, School of Mathematics,
The University of Leeds, UK
[4]Center for Biological Sequence Analysis,
Technical University of Denmark, Lyngby, Denmark

1 Introduction

The problem of protein structure prediction from amino acid sequence (Dill and MacCallum, 2012) can be formulated in the following way. We want to formulate the following probability distribution[1]

$$p(\mathbf{x} \mid N, L, \mathbf{a}) \tag{1}$$

where $\mathbf{x}$ is a vector of dihedral angles (see Figure 1) that specifies a protein's structure. A dihedral angle corresponds to a point on the unit circle; assuming ideal bond angles and bond lengths, a protein can thus be fully parameterized as a sequence of points on the unit circle (Boomsma *et al.*, 2008, Harder *et al.*, 2010). Furthermore, $\mathbf{a}$ is the amino acid sequence – a sequence of symbols chosen from an alphabet with twenty letters – and N and L are conditions that regard nonlocal and local structure, respectively. Local structure concerns protein structure on a local length scale, including α-helices, β-strand, loops and so on. Nonlocal structure concerns features of a more global nature, including hydrogen bonds, amino acid packing in the hydrophobic core and so on. We specify conditioning on protein-like local and nonlocal structure in the following way:

- $p(\mathbf{x} \mid L, N, \mathbf{a})$ is the probability of the dihedral vector $\mathbf{x}$ given the amino acid $\mathbf{a}$ and given that both local (L) and nonlocal (N) structure are required to be protein-like.
- Similarly, $p(\mathbf{x} \mid L, \mathbf{a})$ is the probability of the dihedral vector $\mathbf{x}$ given the amino acid $\mathbf{a}$ and given that the local (L) structure is required to be protein-like; the nonlocal structure (N) is not required to be protein-like in this case.

For our purposes, the configurations that are protein-like with respect to both local and nonlocal structure form a subset of the configurations that are protein-like with respect to local structure alone (see Figure 2). More specifically, the set $\mathbf{x} : p(\mathbf{x} \mid L, N, \mathbf{a}) > 0$ is a subset of $\mathbf{x} : p(\mathbf{x} \mid L, \mathbf{a}) > 0$.

[1] A bold font indicates a vector.

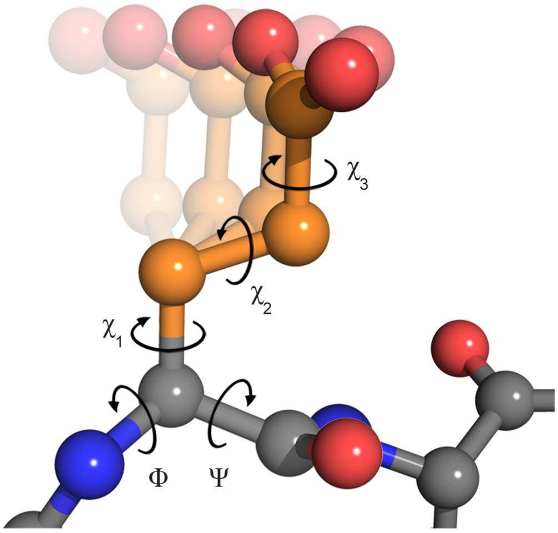

Figure 1: When bond angles and bond lengths are considered as fixed to their ideal values, a vector of dihedral angles is the remaining degree of freedom describing a three-dimensional protein structure. The dihedral angles can be subdivided in backbone and side chain angles, respectively involving (ψ, ϕ, ω) triplets and vectors of χ angles. All angles are illustrated in the figure, with the exception of ω, which is typically close to 180°. The number of χ angles varies between zero and four for the twenty standard amino acids. The figure shows a ball-and-stick representation of a single amino acid, glutamate, which has three χ angles, within a protein. The fading conformations in the background illustrate a rotation around χ_1. Oxygen is shown in red; backbone carbons in grey; side chain carbons in yellow; nitrogen in blue. The figure was made using PyMOL (*http://www.pymol.org*, DeLano Scientific LCC). Figure from (Harder *et al*., 2010).

2 Formulation of the joint model

A direct, practically useful formulation of Equation (1) is intractable. However, in practice, it is possible to formulate the following probability distributions,

$$p(\mathbf{x} \mid L, \mathbf{a}) \tag{2}$$

$$p(\mathbf{e} \mid L, \mathbf{a}) \tag{3}$$

$$p(\mathbf{e} \mid L, N, \mathbf{a}) \tag{4}$$

where $\mathbf{e}$ is some deterministic function $\mathbf{e} = \mathcal{F}(\mathbf{x})$ of $\mathbf{x}$, with $\dim(\mathbf{e}) < \dim(\mathbf{x})$, that provides information on the nonlocal structure associated with $\mathbf{x}$. We refer to the random variable $\mathbf{e}$ as a *coarse grained variable*, while $\mathbf{x}$ is referred to as the *fine grained variable* (Hamelryck *et al*., 2010, Mardia *et al*., 2011, Frellsen *et al*., 2012, Borg *et al*., 2012). The relationship between $\mathbf{x}$ and $\mathbf{e}$ is thus many-to-one. The first probability distribution describes protein structure on a local length scale; the second distribution describes the (hypothetical) nonlocal structure of proteins in the absence of nonlocal interactions such as hydrogen bonds or hydrophobic packing; the final distribution describes the nonlocal structure of actual proteins.

The requested probability distribution given by Equation (1) can be obtained from the probability distributions given by Equations (2,3,4) as follows. First we note that

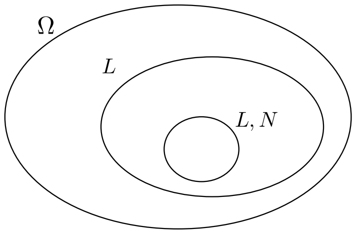

Figure 2: Venn diagram illustrating the relationship between the sets of configurations, or dihedral angle vectors, $\mathbf{x}$ that have a non-zero probability conditioned on L and conditioned on L, N, respectively. Ω is the set of all possible dihedral angle vectors.

$$\begin{aligned} p(\mathbf{x} \mid L, \mathbf{a}) &= p(\mathbf{x}, \mathbf{e} \mid L, \mathbf{a}) \\ &= p(\mathbf{x} \mid \mathbf{e}, L, \mathbf{a}) p(\mathbf{e} \mid L, \mathbf{a}) \end{aligned}$$

where the first step is due to the fact that $\mathbf{e} = \mathcal{F}(\mathbf{x})$, and the second step is a simple application of the product rule of probability. From the above, we derive the following intermediate result

$$p(\mathbf{x} \mid \mathbf{e}, L, \mathbf{a}) = \frac{p(\mathbf{x} \mid L, \mathbf{a})}{p(\mathbf{e} \mid L, \mathbf{a})} \tag{5}$$

Now, similarly

$$\begin{aligned} p(\mathbf{x} \mid L, N, \mathbf{a}) &= p(\mathbf{x}, \mathbf{e} \mid L, N, \mathbf{a}) \\ &= p(\mathbf{x} \mid \mathbf{e}, L, N, \mathbf{a}) p(\mathbf{e} \mid L, N, \mathbf{a}) \\ &= p(\mathbf{x} \mid \mathbf{e}, L, \mathbf{a}) p(\mathbf{e} \mid L, N, \mathbf{a}) \end{aligned} \tag{6}$$

where in the last step we have assumed that

$$p(\mathbf{x} \mid \mathbf{e}, L, N, \mathbf{a}) = p(\mathbf{x} \mid \mathbf{e}, L, \mathbf{a})$$

This assumption of conditional independence is reasonable given the physical interpretation of $\mathbf{e}$, which specifies the nonlocal structure unequivocally, thus rendering the conditioning on N redundant.

Substituting Equation (5) into Equation (6), we obtain our final result

$$p(\mathbf{x} \mid L, N, \mathbf{a}) = \frac{p(\mathbf{e} \mid L, N, \mathbf{a})}{p(\mathbf{e} \mid L, \mathbf{a})} p(\mathbf{x} \mid L, \mathbf{a}) \tag{7}$$

This expression was first reported as the "reference ratio method" by Hamelryck *et al.* in 2010 and developed further in a series of publications (Hamelryck *et al.*, 2010, Mardia *et al.*, 2011, Frellsen *et al.*, 2012, Borg *et al.*, 2012). The reference ratio method solves a long-standing dispute regarding the validity of so-called “potentials of mean force” (Koppensteiner and Sippl, 1998, Thomas and Dill, 1996, Ben-Naim, 1997) – which are widely used in protein structure prediction – that has been ongoing for over twenty years. In this specific case, the generally applicable reference ratio method establishes a joint probability distribution that describes protein structure in atomic detail, in a tractable manner.

3 Implementation and conclusion

The probability distribution $p(\mathbf{x} \mid L, \mathbf{a})$ can be formulated using hidden Markov models whose observed variables are points on a hypertorus or the circle (Boomsma *et al.*, 2008, Harder *et al.*, 2010, Boomsma *et al.*, 2012, Hamelryck *et al.*, 2012). These models have been previously described by us (Boomsma *et al.*, 2008, Harder *et al.*, 2010), and are estimated using a set of experimentally derived protein structures. Combined, these estimated models can be used to formulate the probability distribution given by Equation (3). Hidden Markov models are computationally efficient and tractable; however, their first-order Markov assumption limits their use to the modelling of local structure. This shortcoming can be alleviated by the strategy outlined above.

For $\mathbf{e}$, we choose a low dimensional vector of energy components, concerning hydrogen bonds, electrostatic and hydrophobic interactions, according to a specific physical force field (Irbäck *et al.*, 2009) A probability distribution for $\mathbf{e}$ can be inferred for a given amino acid sequence, and combined with the probability distribution over $\mathbf{x}$ following Equation 7. Protein structure prediction is thus reduced to sampling from the resulting joint probability distribution specified by Equation (7), which can be done using advanced Markov chain Monte Carlo methods (Bottaro *et al.*, 2012, Boomsma *et al.*, 2013, Ferkinghoff-Borg, 2012). I will present some preliminary results obtained using the procedure outlined in this abstract.

Acknowledgements

JV acknowledges funding from the Danish Council for Independent Research (FTP09-066546). JF acknowledges funding from the Danish Council for Strategic Research (NABIIT2106-06-0009). We thank Simon Olsson for comments.

References

Ben-Naim,A. (1997) Statistical potentials extracted from protein structures: Are these meaningful potentials? *J. Chem. Phys.*, **107**, 3698–3706.

Boomsma, W., Frellsen,J. and Hamelryck,T. (2012) Probabilistic models of biomolecular local structure and their applications. In *Bayesian methods in structural; bioinformatics,* (Hamelryck,T., Mardia,K. and Ferkinghoff-Borg,J., eds),. Springer-Verlag Heidelberg, Berlin pp. 233-254.

Boomsma,W., Frellsen,J., Harder,T., Bottaro,S., Johansson,K.E., Tian,P., Stovgaard,K., Andreetta,C., Olsson,S., Valentin,J.B., Antonov,L.D., Christensen,A.S., Borg,M., Jensen,J.H., Lindorff-Larsen,K., Ferkinghoff-Borg,J. and Hamelryck,T. (2013) PHAISTOS: a framework for Markov chain Monte Carlo simulation and inference of protein structure. *J. Comput. Chem.,* **Accepted**.

Boomsma, W., Mardia,K., Taylor,C., Ferkinghoff-Borg,J., Krogh,A. and Hamelryck,T. (2008) A generative, probabilistic model of local protein structure. *Proc. Natl. Acad. USA,* **105**, 8932-8937.

Borg,M., Hamelryck,T. and Ferkinghoff-Borg,J. (2012) On the physical relevance and statistical interpretation of knowledge-based potentials. In *Bayesian methods in structural bioinformatics,* (Hamelryck,T., Mardia,K. and Ferkinghoff-Borg,J., eds),. Springer-Verlag Heidelberg, Berlin pp. 97-124.

Bottaro,S., Boomsma,W., E. Johansson,K., Andreetta,C., Hamelryck,T. and Ferkinghoff-Borg,J. (2012) Subtle Monte Carlo updates in dense molecular systems. *J. Chem. Theory Comput.,* **8**, 695-702.

Dill,K. and MacCallum,J. (2012) The protein-folding problem, 50 years on. *Science,* **338**, 1042-1046.

Ferkinghoff-Borg,J. (2012) Monte Carlo methods for inference in high-dimensional systems. In *Bayesian Methods in Structural Bioinformatics*, (Hamelryck,T., Mardia,K. and Ferkinghoff-Borg,J., eds),. Springer-Verlag Heidelberg, Berlin pp. 49-93.

Frellsen,J., Mardia,K., Borg,M., Ferkinghoff-Borg,J. and Hamelryck,T. (2012) Towards a general probabilistic model of protein structure: the reference ratio method. In *Bayesian methods in structural bioinformatics,* (Hamelryck,T., Mardia,K. and Ferkinghoff-Borg,J., eds),. Springer-Verlag Heidelberg, Berlin pp. 125-134.

Hamelryck,T,. Borg,M., Paluszewski,M., Paulsen,J., Frellsen,J., Andreetta,C., Boomsma,W., Bottaro,S. and Ferkinghoff-Borg,J. (2010) Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS ONE,* **5**, e13714.

Hamelryck,T., Mardia,K. and Ferkinghoff-Borg,J., eds (2012) *Bayesian methods in structural bioinformatics*. Statistics for Biology and Health, Springer-Verlag, Heidelberg, Berlin.

Harder,T., Boomsma, W., Paluszewski,M., Frellsen,J., Johansson,K. and Hamelryck,T. (2010) Betond rotamers: a generative, probabilistic model of side chains in proteins. *BMC Bioinformatics,* **11**, 306.

Irbäck,A., Mitternacht,S. and Mohanty,S. (2009) An effective all-atom potential for proteins. *PMC Biophysics,* **2**,2.

Koppensteiner,W. and Sippl,M. (1998) Knowledge-based potentials-back to the roots. *Biochemistry (Mosc),* **63**, 247-52

Mardia,K., Frellsen,J., Borg,M., Ferkinghoff-Borg,J. and Hamelryck (2011) A statistical view of the reference ratio method. In *LASR2011 - High-throughput sequencing, proteins and statistics,* (Gusnanto,A., Mardia,K. and Fallaize,C., eds), pp. 55-61 Leeds University Press, Leeds, UK.

Thomas,P.D. and Dill,K.A. (1996) Statistical potentials extacted from protein structures: how accurate are they? *J. Mol. Biol.,* **257**, 457-469.