A statistical view on the reference ratio method

Kanti V. Mardia[1], Jes Frellsen[2], Mikael Borg[2],
Jesper Ferkinghoff-Borg[3] and Thomas Hamelryck[2*]

[1]Department of Statistics, University of Leeds.
[2]Bioinformatics Centre, University of Copenhagen.
[3]Department of Electrical Engineering, Technical University of Denmark.

1 Introduction

The recently introduced *reference ratio method* (Hamelryck *et al.*(2010)) allows combining distributions over fine grained variables with distributions over coarse grained variables in a meaningful way. This problem is a major bottleneck in the prediction, simulation and design of protein structure and dynamics. Hamelryck *et al.* introduce the reference ratio method in this context, and show that the method provides a rigorous statistical explanation of the so called *potentials of mean force* (PMFs). These potentials are widely used in protein structure prediction and simulation, but their physical justification is highly disputed (Thomas and Dill(1996); Ben-Naim(1997); Koppensteiner and Sippl(1998)). The reference ratio method clarifies, justifies and extends the scope of these potentials.

As the reference ratio method is of general relevance for statistical purposes, we present the method here in a general statistical setting. Subsequently, we discuss two example applications of the method. First, we present a simple educational example, where the method is applied to independent normal distributions. Secondly, we reinterpret an example originating from Hamelryck *et al.*; in this example, the reference ratio method is used to combine a detailed distribution over the dihedral angles of a protein with a distribution that describes the compactness of the protein. Finally, we outline the relation between the reference ratio method and PMFs.

2 Reference ratio method

We start by introducing the reference ratio method using general statistical terms. Let $f(\boldsymbol{x})$ be the probability density function (pdf) of $\boldsymbol{X}$, which is unknown, but

(i) the pdf $f_1(y)$ of $Y = m(\boldsymbol{X})$ is known for $f(\cdot)$, where $m(\cdot)$ is a specified many-to-one function, and

(ii) the pdf $g(\boldsymbol{x})$ is specified and approximately close to $f(\boldsymbol{x})$, in the sense that $f_2(\cdot|y) \approx g_2(\cdot|y)$ for all practical purposes.

Here, $f_2(\boldsymbol{x}|y)$ denotes the conditional pdf of $\boldsymbol{X}$ given Y for $f(\cdot)$, and $g_2(\boldsymbol{x}|y)$ denotes the corresponding conditional pdf for $g(\cdot)$. Note that these two conditional pdfs are not specified and that their closed form expressions are not necessarily easily expressed. In the work of Hamelryck *et al.*, $\boldsymbol{X}$ is denoted the *fine grained variable* and Y the *coarse grained variable* due to their functional relation.

Now assume that we want to construct a new density $\hat{f}(\boldsymbol{x})$, close to $f(\boldsymbol{x})$, such that

(iii) the marginal pdf of Y for $\hat{f}(\cdot)$ is equal to $f_1(y)$ and

(iv) the conditional pdf of $\boldsymbol{X}$ given $Y = y$ for $\hat{f}(\cdot)$ is equal to $g_2(\boldsymbol{x}|y)$.

In other words $\hat{f}(\boldsymbol{x})$ should have the properties that $\hat{f}_1(y) = f_1(y)$ and $\hat{f}_2(\boldsymbol{x}|y) = g_2(\boldsymbol{x}|y)$, where $\hat{f}_1(y)$ and $\hat{f}_2(\boldsymbol{x}|y)$ respectively denotes the marginal distribution of Y and the conditional distribution of $\boldsymbol{X}$ given Y for $\hat{f}(\cdot)$. It would be straightforward to construct $\hat{f}(\boldsymbol{x})$ if the

conditional pdf $g_2(\boldsymbol{x}|y)$ was known. In particular, generation of samples would be efficient, since we could sample $\tilde{y}$ according to $f_1(\cdot)$ and subsequently sample $\tilde{\boldsymbol{x}}$ according to $g_2(\cdot|y)$, if efficient sampling procedures were available for the two distributions. However, as previously stated $g_2(\boldsymbol{x}|y)$ is generally not known. An approximate solution for sampling could be to approximate the density $g_2(\boldsymbol{x}|y)$ by drawing a large amount of sample according to $g(\boldsymbol{x})$ and retain those with the required value of Y. Obviously, this approach would be intractable for a large sample space.

The solution to the problem was given by Hamelryck *et al.* in the form of a closed form expression for $\hat{f}(\cdot)$. The authors showed that the conditions (iii) and (iv) are satisfied for the pdf given by

$$\hat{f}(\boldsymbol{x}) = \frac{f_1(\hat{y})}{g_1(\hat{y})}\, g(\boldsymbol{x}) \,, \tag{1}$$

where $\hat{y} = m(\boldsymbol{x})$. The construction of the density $\hat{f}(\cdot)$ in the above expression is known as the *reference ratio method* and the corresponding distribution is denoted the *reference ratio distribution* (Hamelryck *et al.*(2010)). It is easy to check that $\hat{f}(\cdot)$ is properly normalized, $\int \hat{f}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$, and that the two conditions (iii) and (iv) are satisfied.

Since Y is a function of $\boldsymbol{X}$, the joint pdf, $g_3(\boldsymbol{x}, y)$, of $(\boldsymbol{X}, Y)$ for $g(\cdot)$ is zero for all values $(\boldsymbol{x}, y')$, where $y' \neq m(\boldsymbol{x})$, and we can write the joint pdf as

$$g_3(\boldsymbol{x}, y) = g(\boldsymbol{x})\, \delta(y - m(\boldsymbol{x})) \,,$$

where $\delta(\cdot)$ is the Dirac delta function. Consequently, we can also express the result from equation (1) as

$$\hat{f}(\boldsymbol{x}) = f_1(\hat{y})\, g_2(\boldsymbol{x}|\hat{y}) \,. \tag{2}$$

Based on this expression, we can recast the result as follows: the pdf $f(\boldsymbol{x})$ is unknown, but its marginal density $f_1(y)$ is known and an approximation, $g_2(\boldsymbol{x}|y)$, of $f_2(\boldsymbol{x}|y)$ is indirectly available through $g(\boldsymbol{x})$ to approximate the density $f(\boldsymbol{x})$. In the following we will present two applications of the reference ratio method.

3 Example with independent normals

The purpose of our first example is purely educational. It is a simple toy example based on independent normal distributions, which simplifies the functional form of the pdfs involved. Let $\boldsymbol{X} = (X_1, X_2)$, where X_1 and X_2 are independent normals with

$$X_1 \sim \mathcal{N}(\mu, 1) \quad \text{and} \quad X_2 \sim \mathcal{N}(0, 1) \,.$$

Accordingly, the pdf of $\boldsymbol{X}$ is given by

$$f(\boldsymbol{x}) = c\, \mathrm{e}^{-\frac{1}{2}(x_1-\mu)^2 - \frac{1}{2}x_2^2} \,,$$

where $\boldsymbol{x} = (x_1, x_2)$ and c is the normalizing constant. For the distribution $g(\boldsymbol{x})$, which is approximately close to $f(\boldsymbol{x})$, let X_1 and X_2 be independently distributed as

$$X_1 \sim \mathcal{N}(0, 1), \quad X_2 \sim \mathcal{N}(0, 1) \,.$$

Consequently the pdf of $\boldsymbol{X}$ is given by

$$g(\boldsymbol{x}) = d\, \mathrm{e}^{-\frac{1}{2}x_1^2 - \frac{1}{2}x_2^2} \,,$$

where d is the normalizing constant. Suppose that $Y = m(\boldsymbol{X}) = X_1$. This means that the marginal pdf of Y for $f(\cdot)$ is

$$f_1(y) = c'\, \mathrm{e}^{-\frac{1}{2}(x_1-\mu)^2} \,,$$

and for $g(\cdot)$ the marginal density is

$$g_1(y) = d' \, \mathrm{e}^{-\frac{1}{2}x_1^2} \, ,$$

where c' and d' are the appropriate normalizing constants. Note that $g(\boldsymbol{x})$ is only a good approximation to $f(\boldsymbol{x})$ for $\mu \simeq 0$, but for both $f(\cdot)$ and $g(\cdot)$ the conditional density of $\boldsymbol{X}$ given Y is the same and equal to the pdf of the normal distribution $\mathcal{N}(0, 1)$.

By applying the ratio method from equation (1), we obtain the expression

$$\hat{f}(\boldsymbol{x}) = \frac{c' \, \mathrm{e}^{-\frac{1}{2}(x_1-\mu)^2} \; d \, \mathrm{e}^{-\frac{1}{2}x_1^2 - \frac{1}{2}x_2^2}}{d' \, \mathrm{e}^{-\frac{1}{2}x_1^2}} = c \, \mathrm{e}^{-\frac{1}{2}(x_1-\mu)^2 - \frac{1}{2}x_2^2} \, .$$

In this example we observed that $\hat{f}(\cdot) = f(\cdot)$, which is expected since the conditional distribution of $\boldsymbol{X}$ given Y is the same for both $f(\cdot)$ and $g(\cdot)$. Accordingly, it is now trivial to check that the marginal distribution of Y for $\hat{f}(\cdot)$ is equal to $f_1(\cdot)$ and that the conditional distribution of $\boldsymbol{X}$ given Y is $g_2(\boldsymbol{x}|y)$, as stated in (iii) and (iv).

Generally, the conditional pdf $g_2(\boldsymbol{x}|y)$ is only assumed to be approximately equal to $f_2(\boldsymbol{x}|y)$, which means that $\hat{f}(\cdot)$ and $f(\cdot)$ are not guaranteed to be equal. In fact, in most relevant applications of the reference ratio method this conditional distribution is unknown for $f(\cdot)$. In next section we will consider such an example.

4 Sampling compact protein structures

A more realistic application of the reference ratio method is given by Hamelryck *et al.*. In this example the method is used to sample compact proteins structures. We will recount the example here using the notation introduced above. The setup is as follows:

(a) Let $f(\boldsymbol{x})$ be an unknown distribution of the dihedral angles $\boldsymbol{x} = \{(\phi_i, \psi_i) \,|\, i = 1, \ldots, n\}$ in a protein with a known sequence of n amino acids.

(b) Let $Y = m(\boldsymbol{X})$ be the radius of gyration (r_g) of the protein, and assume that $f_1(y)$ is a normal distribution with $\mathcal{N}(22\,\text{Å}, 4\,\text{Å}^2)$.

(c) The pdf, $g(\boldsymbol{x})$, of the approximating distribution is given by TorusDBN, which is a probabilistic model of local protein structure (Boomsma *et al.*(2008)).

(d) The marginal density $g_1(y)$ is obtained by sampling from $g(\boldsymbol{x})$, which can be done since TorusDBN is a generative model.

The reference ratio method is applied to construct the density $\hat{f}(\cdot)$, based on the normal distribution over the radius of gyration, $f_1(y)$, the TorusDBN distribution, $g(\boldsymbol{x})$, and the marginal distribution over r_g for TorusDBN, $g_1(y)$. It is important to stress that typical samples generated from TorusDBN, $g(\boldsymbol{x})$, are unfolded and non-compact, while typical samples from $\hat{f}(\boldsymbol{x})$ will be compact as the radius of gyration is controlled by the specified normal distribution. Accordingly, samples from the reference ratio distribution, $\hat{f}(\boldsymbol{x})$, are expected to look more like folded structures than samples from $f(\boldsymbol{x})$.

Hamelryck *et al.* test this setup on the protein ubiquitin, which consists of 76 amino acids. Figure 1 shows the distribution over y (r_g) obtained by sampling from $g(\boldsymbol{x})$ and $\hat{f}(\boldsymbol{x})$, respectively. The figure also shows the normal density $f_1(y)$. We observe that samples from $g(\boldsymbol{x})$ have an average radius of gyration around 27 Å, while samples from $\hat{f}(\boldsymbol{x})$ indeed have a distribution very near $f_1(y)$. As expected, samples from $\hat{f}(\boldsymbol{x})$ are compact, unlike samples from $g(\boldsymbol{x})$. Examples of such samples are shown in figure 2.

A key question here is how can we sample from $\hat{f}(\boldsymbol{x})$ efficiently? As described earlier, we would from a generative point of view use equation (2) directly and generate a sample, $\tilde{\boldsymbol{x}}$, using the two steps:

1. sample $\tilde{y}$ according to $f_1(y)$ and
2. sample $\tilde{\boldsymbol{x}}$ according to $g_2(\boldsymbol{x}|\tilde{y})$.

However, a problem lies in step 2, as there is no efficient way to sample from $g_2(\boldsymbol{x}|y)$; TorusDBN only allows for efficient sampling from $g(\boldsymbol{x})$. One could consider using rejection sampling or the ABC method (Pritchard *et al.*(1999); Beaumont *et al.*(2002); Marjoram *et al.*(2003)) for step 2, but both methods would be very inefficient. Hamelryck *et al.* (2010) have given a highly efficient method, which does not (in principle) involve any approximations. The idea is to use the Metropolis-Hastings algorithm with $g(\boldsymbol{x})$ as proposal distribution and $\hat{f}(\boldsymbol{x})$ as target distribution. In this case, the probability of accepting a proposed value $\boldsymbol{x'}$ given a previous values $\boldsymbol{x}$ becomes

$$\alpha(\boldsymbol{x'}|\boldsymbol{x}) = \min\left(1, \frac{f_1(y')g(\boldsymbol{x'})/g_1(y')}{f_1(y)g(\boldsymbol{x})/g_1(y)} \frac{g(\boldsymbol{x})}{g(\boldsymbol{x'})}\right) = \min\left(1, \frac{f_1(y')}{f_1(y)} \frac{g_1(y)}{g_1(y')}\right) , \tag{3}$$

where $y = m(\boldsymbol{x})$ and $y' = m(\boldsymbol{x'})$. In practice, the proposal distribution in the MCMC algorithm would only change a randomly chosen consecutive subsequence of $\boldsymbol{X}$ using TorusDBN (see supporting information of Boomsma *et al.* (2008) for details), as this leads to a higher acceptance rate. It can be shown that the acceptance probability in this case also is given by equation (3).

5 The reference ratio method explains PMFs

Methods for predicting the structure of proteins rely on an energy function or probability distribution that describes the space of possible conformations. One approach to constructing such energies or distributions is to estimate them from a set of experimental determined protein structures. In this case they are called *knowledge based potentials*.

A subclass of the knowledge based potentials are based on probability distributions over pairwise distances in proteins. These are called *potentials of mean force* (PMFs) and are loosely based on an analogy with the statistical physics of liquids (Ben-Naim(1997); Koppensteiner and Sippl(1998)). The potential of mean force, $W(\boldsymbol{r})$, associated with a set of pairwise distances $\boldsymbol{r}$ is given by an expression of the form

$$W(\boldsymbol{r}) \propto -\log \frac{f_1(\boldsymbol{r})}{g_1(\boldsymbol{r})} ,$$

where $f_1(\boldsymbol{r})$ is a pdf estimated from a database of known protein structure, and $g_1(\boldsymbol{r})$ is the pdf of $\boldsymbol{r}$ for a so-called *reference state*. The reference state is typically defined based on physical considerations. The pdf $f_1(\boldsymbol{r})$ is constructed by assuming that the individual pairwise distances are conditionally independent, which constitutes a crude approximation. In practice, the potential of mean force is combined with an additional energy function, that is concerned with the local structure of proteins. This additional energy term is typically brought in via sampling from a *fragment library* (Simons *et al.*(1997)) – a set of short fragments derived from experimental protein structures – or any other sampling method that generates protein-like conformations. From a statistical point of view, this means that the samples are generated according to the pdf

$$\hat{f}(\boldsymbol{x}) \propto \frac{f_1(\boldsymbol{r})}{g_1(\boldsymbol{r})} g(\boldsymbol{x}) , \tag{4}$$

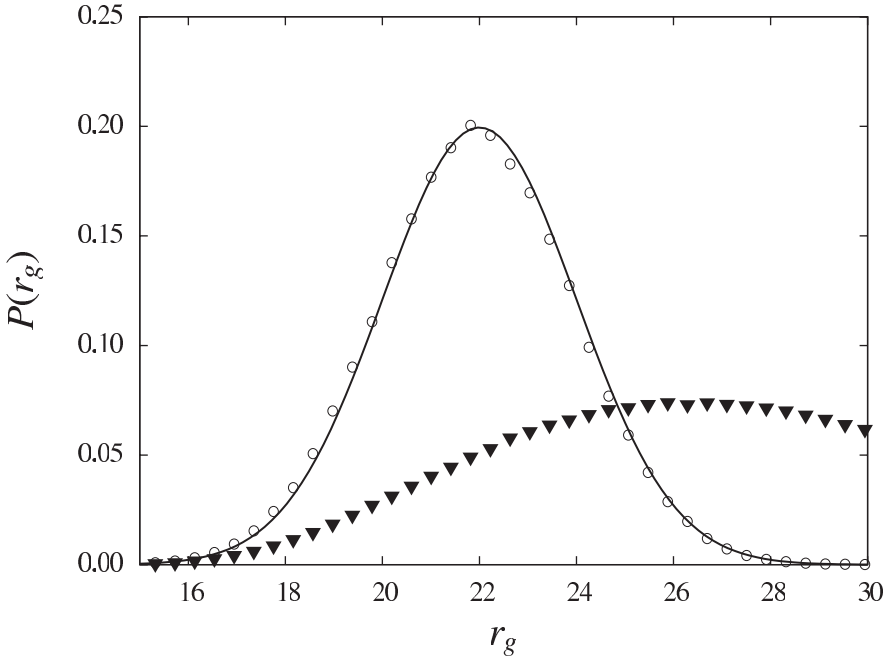

Figure 1: The reference ratio method applied to sampling protein structures with a specified distribution over the radius of gyration (r_g). The distribution over r_g for samples from TorusDBN, $g(\boldsymbol{x})$, is shown as triangles, while the r_g-distribution for samples from the ratio distribution, $\hat{f}(\boldsymbol{x})$, is shown as circles. The pdf, $f_1(y)$, for the desired distribution normal distribution over r_g is shown as a solid line, $\mathcal{N}(22\,\text{Å}, 4\,\text{Å}^2)$. The samples are produced using the amino acid sequence of ubiquitin. The figure is adapted from figure 3 in (Hamelryck *et al.*(2010)).

where $\boldsymbol{x}$ are the dihedral angles in the protein, $\boldsymbol{r}$ are the pairwise distances implied by $\boldsymbol{x}$, and $g(\boldsymbol{x})$ is the pdf of the dihedral angles embodied in the sampling method.

In this formulation, it can be seen that PMFs are justified by the reference ratio method; their functional form arises from the combination of the sampling method (which concerns the fine grained variable) with the pairwise distance information (which concerns the coarse grained variable). This interpretation of PMFs also provides some surprising new insights. First, $g_1(\boldsymbol{r})$ is uniquely defined by $g(\boldsymbol{x})$, and does not require any external physical considerations. Second, if the three involved probability distributions are properly defined, the PMF approach is entirely rigorous and statistically well justified. Third, the PMF approach generalizes beyond pairwise distances to arbitrary coarse grained variables. In conclusion, the reference ratio method settles a dispute over the validity of PMFs that has been going on for more than twenty years, and opens the way to efficient and well-justified probabilistic models of protein structure.

6 Acknowledgements

The authors acknowledge funding by the Danish Program Commission on Nanoscience, Biotechnology and IT (NaBiIT, project: Simulating proteins on a millisecond time-scale, 2106-06-0009).

References

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**(4), 2025–2035.

Ben-Naim, A. (1997). Statistical potentials extracted from protein structures: Are these mean-

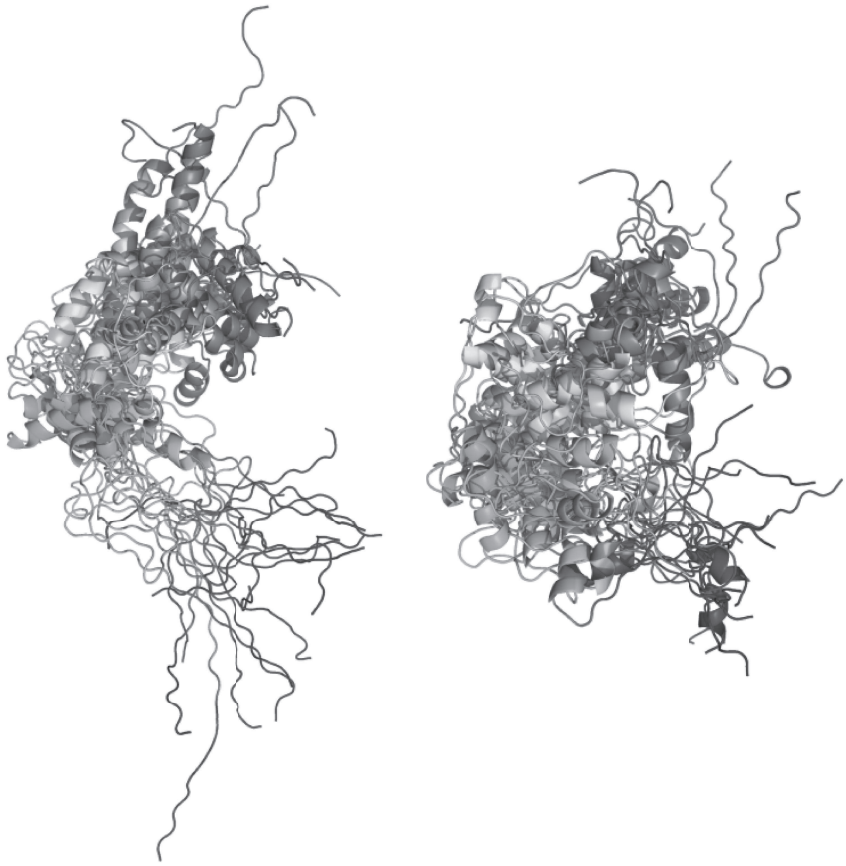

Figure 2: The reference ratio method applied to sample compact protein structures. The left picture shows twenty typical, non-compact samples obtained using TorusDBN alone, $g(\boldsymbol{x})$. The right picture shows twenty typical, compact samples from the reference ratio distribution, $\hat{f}(\boldsymbol{x})$, which is obtained by combining TorusDBN, $g(\boldsymbol{x})$, with a suitable normal distribution over the radius of gyration, $f_1(y) = \mathcal{N}(22\,\text{Å}, 4\,\text{Å}^2)$. We used the amino acid sequence of ubiquitin.

ingful potentials? *J Chem Phys*, **107**(9), 3698–3706.

Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A., and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci U S A*, **105**(26), 8932–8937.

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**(11), e13714.

Koppensteiner, W. A. and Sippl, M. J. (1998). Knowledge-based potentials–back to the roots. *Biochemistry Mosc*, **63**(3), 247–252.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A*, **100**(26), 15324–15328.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol*, **16**(12), 1791–1798.

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, **268**(1), 209–225.

Thomas, P. D. and Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol*, **257**(2), 457–469.