ABC methods for model choice in Gibbs random fields Application to protein 3D structure prediction

by

Aude Grelaud, Christian P. Robert, Jean-Michel Marin, François Rodolphe and Jean-François Taly

Research Report No. 18

September 2008

Statistics for Systems Biology Group

Jouy-en-Josas/Paris/Evry, France

http://genome.jouy.inra.fr/ssb/

ABC methods for model choice in Gibbs random fields
Application to protein 3D structure prediction

Aude Grelaud[⋆,†,‡], Christian P. Robert[⋆,†], Jean-Michel Marin[§,†],
François Rodolphe[‡] and Jean-François Taly[‡]

[‡] INRA, unité MIG, 78350 Jouy en Josas, France

[⋆] CEREMADE, Université Paris Dauphine, 75775 Paris cedex 16, France

[†] CREST, INSEE, 92145 Malakoff cedex, France

[§] INRIA Saclay, Projet select, Université Paris-Sud, 91400 Orsay, France

Abstract

Gibbs random fields are polymorphous statistical models that can be used to analyse different types of dependence, in particular for spatially correlated data. However, when those models are faced with the challenge of selecting a dependence structure from many, the use of standard model choice methods is hampered by the unavailability of the normalising constant in the Gibbs likelihood. In particular, from a Bayesian perspective, the computation of the posterior probabilities of the models under competition requires special likelihood-free simulation techniques like the Approximate Bayesian Computation (ABC) algorithm that is intensively used in population Genetics. We show in this paper how to implement an ABC algorithm geared towards model choice in the general setting of Gibbs random fields, demonstrating in particular that there exists a sufficient statistic across models. The accuracy of the approximation to the posterior probabilities can be further improved by importance sampling on the distribution of the models. The practical aspects of the method are detailed through two applications, the test of an iid Bernoulli model versus a first-order Markov chain, and the choice of a folding structure for a protein of *Thermotoga maritima* implicated into signal transduction processes.

Keywords: Approximate Bayesian computation, model choice, Markov random fields, Bayes factor, protein folding.

1 Introduction

Gibbs random fields are probabilistic models associated with the likelihood function

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}} \exp\{\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})\}\,, \tag{1}$$

where $\mathbf{x}$ is a vector of dimension n taking values over $\mathcal{X}$ (possibly a lattice), $S(\cdot)$ is the potential function defining the random field, taking values in $\mathbb{R}^p$, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the associated parameter, and $Z_{\boldsymbol{\theta}}$ is the corresponding normalising constant. When considering model selection within this class of Gibbs models, the primary difficulty to address is the unavailability of the normalising constant $Z_{\boldsymbol{\theta}}$. In most realistic settings, the summation

$$Z_{\boldsymbol{\theta}} = \sum_{\mathbf{x}\in\mathcal{X}} \exp\{\boldsymbol{\theta}^{\mathrm{T}} S(\mathbf{x})\}$$

involves too many terms to be manageable and numerical approximations like path sampling [2], pseudo likelihood or those based on an auxiliary variable [7] are not necessarily available.

Selecting a model with potential S_0 taking values in $\mathbb{R}^{p_0}$ versus a model with potential S_1 taking values in $\mathbb{R}^{p_1}$ depends on the Bayes factor corresponding to the priors π_0 and π_1 on each parameter space

$$BF_{m_0/m_1}(\mathbf{x}) = \int \exp\{\boldsymbol{\theta}_0^{\mathrm{T}} S_0(\mathbf{x})\}/Z_{\boldsymbol{\theta}_0,0}\pi_0(\mathrm{d}\boldsymbol{\theta}_0) \Big/ \int \exp\{\boldsymbol{\theta}_1^{\mathrm{T}} S_1(\mathbf{x})\}/Z_{\boldsymbol{\theta}_1,1}\pi_1(\mathrm{d}\boldsymbol{\theta}_1)$$

but this is not directly achievable. One faces the same difficulty with the posterior probabilities of the models since they depend on those unknown constants. To properly approximate those posterior quantities, it is thus necessary to use likelihood-free techniques such as ABC [8] and we show in this paper how ABC is naturally tuned for this purpose by providing a direct estimator of the Bayes factor.

A special but important case of Gibbs random fields where model choice is particularly crucial is found with Markov random fields (MRF). Those models are special cases of Gibbs random fields used to model the dependency within spatially correlated data, with applications in epidemiology [3] and image analysis [4], among others. In this setting, the potential function S takes values in $\mathbb{R}$ and is associated with a neighbourhood structure denoted by $i \sim i'$. It means that the conditionnal density on x_i only depends on the $x_{i'}$ such that i and i' are neighbours. For instance, the potential function of a Potts model is

$$S(\mathbf{x}) = \sum_{i'\sim i} \mathbb{I}_{\{x_i = x_{i'}\}}\,,$$

where $\sum_{i' \sim i}$ indicates that the summation is taken over all the neighbours pairs. In that case, θ is a scalar. The potential function therefore monitors the number of likewise neighbours over $\mathcal{X}$.

For a fixed neighbourhood, the unavailability of Z_θ complicates inference on the scale parameter θ, but the difficulty is increased manifold when several neighbourhood structures are under comparison on the basis of an observation from (1). In this paper, we will consider the toy example of an iid sequence [with trivial neighbourhood structure] tested against a Markov chain model [with nearest neighbour structure] as well as a biophysical example aimed at selecting a protein 3D structure.

2 Methods

2.1 ABC

As noted above, when the likelihood is not available in closed form, there exist likelihood-free methods that overcome the difficulty faced by standard simulation techniques via a basic acceptance-rejection algorithm. The algorithm on which the ABC method [introduced by [8] and expanded in [1] and [6]] is based can be briefly described as follows: given a dataset $\mathbf{x}^0$ associated with the sampling distribution $f(\cdot|\theta)$, and under a prior distribution $\pi(\theta)$ on the parameter θ, this method generates a parameter value from the posterior distribution $\pi(\theta|\mathbf{x}^0) \propto \pi(\theta) f(\mathbf{x}^0|\theta)$ by simulating a value θ^* from the prior, $\theta^* \sim \pi(\cdot)$, then a value x^* from the sampling distribution $\mathbf{x}^* \sim f(\cdot|\theta^*)$ until $\mathbf{x}^*$ is equal to the observed dataset $\mathbf{x}^0$. The rejection algorithm thus reads as

---

Algorithm 1. Exact rejection algorithm:

- *Generate* θ^* *from the prior* π.
- *Generate* $\mathbf{x}^*$ *from the model* $f(\cdot|\theta^*)$.
- *Accept* θ^* *if* $\mathbf{x}^* = \mathbf{x}^0$.

This solution is not approximative in that the output is truly simulated from the posterior distribution $\pi(\theta|\mathbf{x}^0)$. In many settings, including those with continuous observations $\mathbf{x}^0$, it is however impractical or impossible to wait for $\mathbf{x}^* = \mathbf{x}^0$ to occur and the approximative solution is to introduce a tolerance in the test, namely to accept θ^* if simulated data and observed data are close enough, in the sense of a distance $\rho(.)$, given a fixed tolerance level ϵ.
The ϵ-tolerance rejection algorithm is then

Algorithm 2. ϵ-tolerance rejection algorithm:

- *Generate* θ^* *from the prior* π.
- *Generate* $\mathbf{x}^*$ *from the model* $f(\cdot|\theta^*)$.
- *Accept* θ^* *if* $\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon$.

This approach is obviously approximative when $\epsilon \neq 0$ and it amounts to simulating from the prior when $\epsilon \to \infty$. The output from the algorithm 2 is thus associated with the distribution $\pi(\theta|\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon)$. The choice of ϵ is therefore paramount for good performances of the method. If ϵ is too large, the approximation is poor while, if ϵ is sufficiently small, $\pi(\theta|\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon)$ is a good approximation of $\pi(\theta|\mathbf{x}^0)$ but the acceptance probability may be too low to be practical. It is therefore customary to pick ϵ as an empirical quantile of $\rho(\mathbf{x}^*, \mathbf{x}^0)$ when $\mathbf{x}^*$ is simulated from the marginal distribution, and the choice often is the corresponding 1% quantile (see [1]).

The data $\mathbf{x}^0$ usually being of a large dimension, another level of approximation is enforced within the ABC algorithm, by replacing the distance $\rho(\mathbf{x}^*, \mathbf{x}^0)$ with a corresponding distance between summary statistics $\rho(S(\mathbf{x}^*), S(\mathbf{x}^0))$ [1]. It is straightforward to see that, when $S(.)$ is a sufficient statistic, this step has no impact on the approximation but it is rarely the case that a sufficient statistic of low dimension is available when implementing ABC. As it occurs, the setting of model choice among Gibbs random fields allows for such a beneficial structure, as will be seen below. In the general case, the output of the ABC algorithm is therefore a simulation from the distribution $\pi(\theta|\rho(S(\mathbf{x}^*), S(\mathbf{x}^0)) < \epsilon)$. The algorithm reads as follows:

Algorithm 3. ABC algorithm:

- *Generate* θ^* *from the prior* π.
- *Generate* $\mathbf{x}^*$ *from the model* $f(\cdot|\theta^*)$.
- *Compute the distance* $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.
- *Accept* θ^* *if* $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$.

2.2 Model choice via ABC

In a model choice perspective, we face M Gibbs random fields in competition, each one being associated with a potential function S_m $(0 \leq m \leq M-1)$, i.e. with corresponding likelihood

$$f_m(\mathbf{x}|\theta_m) = \exp\left\{\theta_m^{\mathrm{T}} S_m(\mathbf{x})\right\} / Z_{\theta_m,m}\,,$$

where $\theta_m \in \Theta_m$ and $Z_{\theta_m,m}$ is the unknown normalising constant. Typically, the choice is between M neighbourhood relations $i \overset{m}{\sim} i'$ $(0 \leq m \leq M-1)$ with $S_m(\mathbf{x}) = \sum_{i \overset{m}{\sim} i'} \mathbb{I}_{\{x_i = x_{i'}\}}$.

From a Bayesian perspective, the choice between those models is driven by the posterior probabilities of the models. Namely, if we consider an extended parameter that includes the model index $\mathcal{M}$ defined on the parameter space $\Theta = \cup_{m=0}^{M-1}\{m\} \times \Theta_m$, we can define a prior distribution on the model index $\pi(\mathcal{M} = m)$ as well as a prior distribution on the parameter conditional on the value m of the model index, $\pi_m(\theta_m)$, defined on the parameter space Θ_m. The computational target is thus the model posterior probability

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) \propto \int_{\Theta_m} f_m(\mathbf{x}|\theta_m)\pi_m(\theta_m)\,\mathrm{d}\theta_m\,\pi(\mathcal{M} = m)\,,$$

i.e. the marginal of the posterior distribution on $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$ given $\mathbf{x}$. Therefore, if $S(.)$ is a sufficient statistics for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$,

$$\mathbb{P}(\mathcal{M} = m|\mathbf{x}) = \mathbb{P}(\mathcal{M} = m|S(\mathbf{x}))\,.$$

Each model has its own sufficient statistic $S_m(\cdot)$. Then, for each model, the vector of statistics $S(\cdot) = (S_0(\cdot), \ldots, S_{M-1}(\cdot))$ is obviously sufficient (since it includes the sufficient statistic of each model). Moreover, the structure of the Gibbs random field allows for a specific factorisation of the distribution $f_m(\mathbf{x}|\theta_m)$. Indeed, the distribution of $\mathbf{x}$ in model m factorises as

$$\begin{aligned} f_m(\mathbf{x}|\theta_m) &= f_m^1(\mathbf{x}|S(\mathbf{x}))f_m^2(S(\mathbf{x})|\theta_m) \\ &= \frac{1}{n(S(\mathbf{x}))} f_m^2(S(\mathbf{x})|\theta_m) \end{aligned}$$

where $f_m^2(S(\mathbf{x})|\theta_m)$ is the distribution of $S(\mathbf{x})$ within model m [not to be confused with the distribution of $S_m(\mathbf{x})$] and

$$n(S(\mathbf{x})) = \sharp\left\{\tilde{\mathbf{x}} \in \mathcal{X} : S(\tilde{\mathbf{x}}) = S(\mathbf{x})\right\}$$

is the cardinal of the set of elements of $\mathcal{X}$ with the same sufficient statistic, which does not depend on m (the support of f_m is constant with m). The statistic $S(.)$ is therefore also sufficient for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$.

That the concatenation of the sufficient statistics of each model is also a sufficient statistic for the joint parameters $(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})$ is obviously a property that is specific to Gibbs random field models.

For Gibbs random fields models, it is then possible to apply the above ABC algorithm in order to produce an approximation with tolerance factor ϵ:

Algorithm 4. ABC algorithm for model choice (ABC-MC):

- *Generate* m^* *from the prior* $\pi(\mathcal{M} = m)$.
- *Generate* $\theta^*_{m^*}$ *from the prior* $\pi_{m^*}(\cdot)$.
- *Generate* $\mathbf{x}^*$ *from the model* $f_{m^*}(\cdot|\theta^*_{m^*})$.
- *Compute the distance* $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.
- *Accept* $(\theta^*_{m^*}, m^*)$ *if* $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$.

For the same reason as above, this algorithm results in an approximate generation from the joint posterior distribution

$$\pi\left\{(\mathcal{M}, \theta_0, \ldots, \theta_{M-1})|\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon\right\} .$$

When it is possible to achieve $\epsilon = 0$, the algorithm is exact since S is a sufficient statistic. We have thus derived a likelihood-free method to handle model choice.

Once a sample of N values of $(\theta^{i*}_{m^{i*}}, m^{i*})$ $(1 \leq i \leq N)$ is generated from this algorithm, a standard Monte Carlo approximation of the posterior probabilities is provided by the empirical frequencies of visits to the model, namely

$$\widehat{\mathbb{P}}(\mathcal{M} = m|\mathbf{x}^0) = \sharp\{m^{i*} = m\}/N\,,$$

where $\sharp\{m^{i*} = m\}$ denotes the number of simulated m^{i*}'s equal to m.

Correlatively, the Bayes factor associated with the evidence provided by the data $\mathbf{x}^0$ in favour of model m_0 relative to model m_1 is defined by

$$\begin{aligned} BF_{m_0/m_1}(\mathbf{x}^0) &= \frac{\mathbb{P}(\mathcal{M} = m_0|\mathbf{x}^0)}{\mathbb{P}(\mathcal{M} = m_1|\mathbf{x}^0)} \frac{\pi(\mathcal{M} = m_1)}{\pi(\mathcal{M} = m_0)} \\ &= \frac{\int f_{m_0}(\mathbf{x}^0|\theta_0)\pi_0(\theta_0)\pi(\mathcal{M} = m_0)\mathrm{d}\theta_0}{\int f_{m_1}(\mathbf{x}^0|\theta_1)\pi_1(\theta_1)\pi(\mathcal{M} = m_1)\mathrm{d}\theta_1} \frac{\pi(\mathcal{M} = m_1)}{\pi(\mathcal{M} = m_0)}\,. \end{aligned}$$

The previous estimates of the posterior probabilities can then be plugged-in to approximate the above Bayes factor by

$$\begin{aligned}\overline{BF}_{m_0/m_1}(\mathbf{x}^0) &= \frac{\hat{\mathbb{P}}(\mathcal{M}=m_0|\mathbf{x}^0)}{\hat{\mathbb{P}}(\mathcal{M}=m_1|\mathbf{x}^0)} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)} \\ &= \frac{\sharp\{m^{i*}=m_0\}}{\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}\,,\end{aligned}$$

but this estimate is only defined when $\sharp\{m^{i*}=m_1\} \neq 0$. To bypass this difficulty, the substitute

$$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{1+\sharp\{m^{i*}=m_0\}}{1+\sharp\{m^{i*}=m_1\}} \times \frac{\pi(\mathcal{M}=m_1)}{\pi(\mathcal{M}=m_0)}$$

is particularly interesting because we can evaluate its bias. (Note that there does not exist an unbiased estimator of $BF_{m_0/m_1}(\mathbf{x}^0)$ based on the m^{i*}'s.) Indeed, assuming without loss of generality that $\pi(\mathcal{M}=m_1) = \pi(\mathcal{M}=m_0)$, if we set $N_0 = \sharp\{m^{i*}=m_0\}$, $N_1 = \sharp\{m^{i*}=m_1\}$ and conditionally on $N = N_0 + N_1$, N_1 is a binomial $\mathcal{B}(N,p)$ rv with probability $p = (1+BF_{m_0/m_1}(\mathbf{x}^0))^{-1}$. It is then straightforward to establish that (see Appendix)

$$\mathbb{E}\left[\frac{N_0+1}{N_1+1}\middle| N\right] = BF_{m_0/m_1}(\mathbf{x}^0) + \frac{1}{p(N+1)} - \frac{N+2}{p(N+1)}(1-p)^{N+1}\,.$$

The bias in the estimator $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ is thus $\{1-(N+2)(1-p)^{N+1}\}/(N+1)p$, which goes to zero as N goes to infinity.

2.3 Two step ABC

The above estimator $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ is rather unstable (i.e. suffers from a large variance) when $BF_{m_0/m_1}(\mathbf{x}^0)$ is very large since, when $\mathbb{P}(\mathcal{M}=m_1|\mathbf{x}^0)$ is very small, $\sharp\{m^{i*}=m_1\}$ is most often equal to zero. This difficulty can be bypassed by a reweighting scheme. If the choice of m^* in the ABC algorithm is driven by the probability distribution $\mathbb{P}(\mathcal{M}=m_1) = \varrho = 1-\mathbb{P}(\mathcal{M}=m_0)$ rather than by $\pi(\mathcal{M}=m_1) = 1-\pi(\mathcal{M}=m_0)$, the value of $\sharp\{m^{i*}=m_1\}$ can be increased and later corrected by considering instead

$$\widetilde{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{1+\sharp\{m^{i*}=m_0\}}{1+\sharp\{m^{i*}=m_1\}} \times \frac{\varrho}{1-\varrho}\,.$$

Therefore, if a first run of the ABC algorithm exhibits a very large value of $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$, the estimate $\widetilde{BF}_{m_0/m_1}(\mathbf{x}^0)$ produced by a second run with

$$\varrho \propto 1 \Big/ \hat{\mathbb{P}}(\mathcal{M}=m_1|\mathbf{x}^0)$$

will be more stable than the original $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$. In the most extreme cases when no m^{i*} is ever equal to m_1, this corrective second is unlikely to bring much stabilisation, though. Note, however, that, from a practical point of view, obtaining a poor evaluation of $BF_{m_0/m_1}(\mathbf{x}^0)$ when the Bayes factor is very small (or very large) has limited consequences since the poor approximation also leads to the same conclusion about the choice of model m_0.

3 Results

3.1 Toy example

Our first example compares an iid Bernoulli model with a two-state first-order Markov chain. Both models are special cases of MRF and, furthermore, the normalising constant $Z_{\theta_m,m}$ can be computed in closed form, as well as the posterior probabilities of both models. We thus consider a sequence $\mathbf{x} = (x_1, .., x_n)$ of n binary variables ($n = 100$). Under model $\mathcal{M} = 0$, the MRF representation of the Bernoulli distribution $\mathcal{B}(1/\{1 + exp(-\theta_0)\})$ is

$$f_0(\mathbf{x}|\theta_0) = \exp\left(\theta_0 \sum_{i=1}^{n} \mathbb{I}_{\{x_i=1\}}\right) \big/ \{1 + exp(\theta_0)\}^n \,,$$

associated with the sufficient statistic $S_0(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}_{\{x_i=1\}}$ and the normalising constant $Z_{\theta_0,0} = (1 + e^{\theta_0})^n$. Under a uniform prior $\theta_0 \sim \mathcal{U}(-5, 5)$, the bound ± 5 being the phase transition value for the Gibbs random field, the posterior probability of this model is available since the marginal when $S_0(\mathbf{x}) = s_0$ ($s_0 \neq 0$) is given by

$$\frac{1}{10} \sum_{k=0}^{s_0-1} \binom{s_0 - 1}{k} \frac{(-1)^{s_0-1-k}}{n - 1 - k} \left[(1 + e^5)^{k-n+1} - (1 + e^{-5})^{k-n+1}\right] ,$$

by a straightforward rational fraction integration.

Model $\mathcal{M} = 1$ is chosen as a Markov chain (hence a particular MRF in dimension one with i and i' being neighbours if $|i - i'| = 1$) with the special feature that the probability to remain within the same state is constant over both states, namely

$$\mathbb{P}(x_{i+1} = x_i|x_i) = \exp(\theta_1)\big/\{1 + \exp(\theta_1)\} \,.$$

We assume a uniform distribution on x_1 and the likelihood function for this model is thus

$$f_1(\mathbf{x}|\theta_1) = \frac{1}{2} \exp\left(\theta_1 \sum_{i=2}^{n} \mathbb{I}_{\{x_i=x_{i-1}\}}\right) \big/ \{1 + \exp(\theta_1)\}^{n-1} \,,$$

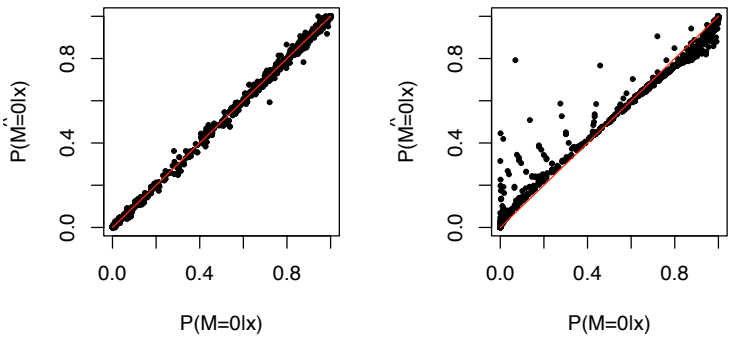

Figure 1: *(left)* Comparison of the true $\mathbb{P}(\mathcal{M} = 0|\mathbf{x}^0)$ with $\widehat{\mathbb{P}}(\mathcal{M} = 0|\mathbf{x}^0)$ over $2,000$ simulated sequences and 4.10^6 proposals from the prior. The red line is the diagonal. *(right)* Same comparison when using a tolerance ϵ corresponding to the 1% quantile on the distances.

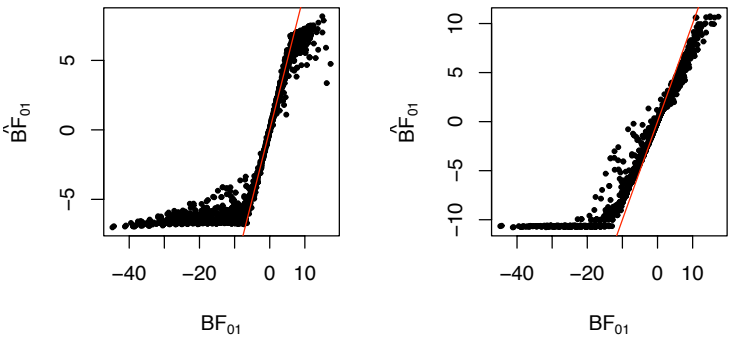

Figure 2: *(left)* Comparison of the true $BF_{m_0/m_1}(\mathbf{x}^0)$ with $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ (in logarithmic scales) over $2,000$ simulated sequences and 4.10^6 proposals from the prior. The red line is the diagonal. *(right)* Same comparison when using a tolerance corresponding to the 1% quantile on the distances.

with $S_1(\mathbf{x}) = \sum_{i=2}^{n} \mathbb{I}_{\{x_i=x_{i-1}\}}$ being the sufficient statistic in that case. Under a uniform prior $\theta_1 \sim \mathcal{U}(0,6)$, where the bound 6 is the phase transition value, the posterior probability of this model is once again available, the likelihood being of the same form as when $\mathcal{M} = 0$.

We are therefore in a position to evaluate the ABC approximations of the model posterior probabilities and of the Bayes factor against the exact values. For this purpose, we simulated 1000 datasets $\mathbf{x}^0$ under each model, using parameters simulated from the priors and computed the exact posterior probabilities and the Bayes factors in both cases. Here, the vector of summary statistics is composed of the sufficient statistic of each model, $S(.) = (S_0(.), S_1(.))$, and we use an euclidian distance.

For each of those datasets $\mathbf{x}^0$, the ABC-MC algorithm was run for 4.10^6 loops, meaning that 4.10^6 sets $(m^*, \theta^*_{m^*}, \mathbf{x}^*)$ were simulated and a random number of those were accepted when $S(\mathbf{x}^*) = S(\mathbf{x}^0)$. (In the worst case scenarios, the number of acceptances was 12!)

As shown on the left graph of Figure 1, the fit of posterior probabilities is very good for all values of $\mathbb{P}(\mathcal{M} = 0|\mathbf{x}^0)$. When we introduce a tolerance ϵ equal to the 1% quantile of the distance, the results are very similar when $\mathbb{P}(\mathcal{M} = 0|\mathbf{x}^0)$ is close to 0, 1 or 0.5, and we observe a slight difference between these values. We also evaluated the approximation of the Bayes factor (and of the subsequent model choice) against the exact Bayes factor. As clearly pictured on the left graph of Figure 2, the fit is very good in the exact case ($\epsilon = 0$), the poorest fits occurring in the limiting cases when the Bayes factor is either very large or very small and thus when the model choice is not an issue, as noted above. In the central zone when $\log BF_{m_0/m_1}(\mathbf{x}^0)$ is close to 0, the difference is indeed quite small, the few diverging cases being due to occurrences of very small acceptance rates. Once more, using a tolerance ϵ equal to the 1% quantile does not bring much difference in the output, the approximative Bayes factor being slightly less discriminative in that case (since the slope of the cloud is less than the even slope of the diagonal).

Given that using the tolerance version allows for more simulations to be used in the Bayes factor approximation, we thus recommend using this approach.

3.2 Application to protein 3D structure prediction

In Biophysics, the knowledge of the tridimensional structure, called *folding,* of a protein provides important information about the protein, including its function.

There exist two kinds of methods that are used to find the 3D structure of a protein. Experimental methods like spectroscopy provide accurate descriptions of structures, but these methods are time consuming, expensive, and sometimes unsuccessful. Alternatively, computational methods

have become more and more successful in predicting 3D structures. These latter methods mostly rely on homology (two sequences are said to be homolog if they share a common ancestor). If the protein under study, hereafter called the query protein, displays a sufficiently high sequence similarity with a protein of known structure, both are considered as homologous with similar structures. Then, a prediction based on the structure of its homolog can be built.

As structures are more conserved over time than sequences, methods based on *threading* have been developed when sequence similarity is too low to assess homology with sufficient certainty. Threading consists in trying to fold the query protein onto all known structures; a fitting criterion is computed for each proposal and the structures displaying sufficiently high criterion values, if any, are chosen. The result is a list of alignments of the query sequence on different structures. It may happen that some of those criterion values are close enough to make a decision difficult.

From a statistical perspective, each structure can be represented by a graph such that a node of this graph is one amino-acid of the protein and an edge between two nodes of the graph indicates that both amino-acids are in close contact in the folded protein. This graph can thus be associated with a Markov random field (1) if labels are allocated to each node. When several structures are proposed by a threading method, algorithm 4 is then available to select the most likely structure. In the example we discuss in this section, labels are associated with the hydrophobic properties of the amino-acids, i.e. they are classified as being either hydrophobic or hydrophilic, based on Table 1. Since, for a globular protein in its native folding, hydrophobic amino-acids are mostly buried inside the structure, and hydrophilic ones are exposed to water, using hydrophobia as the clustering factor does make sense.

Hydrophilic	Hydrophobic
K E R D Q N P H S T G	A Y M W F V L I C

Table 1: Classification of amino acids into hydrophilic and hydrophobic groups.

The dataset is built around a globular protein sequence of known structure (this one will be called the native structure) **1tqgA** corresponding to a protein of *Thermotoga maritima* involved into signal transduction processes. We used FROST [5], a software dedicated to threading, and MODELLER [9] to build candidate structures. (All proteins we use are available in the Protein Data Bank at http://www.rcsb.org/pdb/home/home.do). These candidates thus correspond to the same sequence (the query sequence) folded on different structures. We used two criteria to qualify the similarity between the candidates and the query : the percentage of identity between query and candidate sequences (computed by FROST) and the similarity of the topologies of candidate

structures with the native structure (assessed with the TM-score, 11). Among those, we selected candidates covering the whole spectrum of predictions that can be generated by protein threading, from good to very poor as [10] and described in Table 2.

	% seq . Id.	TM-score	FROST score
1i5nA (**ST1**)	32	0.86	75.3
1ls1A1 (**ST2**)	5	0.42	8.9
1jr8A (**ST3**)	4	0.24	8.9
1s7oA (**DT**)	10	0.08	7.8

Table 2: Summary of the characteristics of our dataset. *% seq . Id.*: percentage of identity with the query sequence. *TM-score.*: similarity between a predicted structure and the native structure. A score larger than 0.4 implies a structural relationship between 2 structures and a score less than 0.17 means that the prediction is nothing more than a random selection from the PDB library. *FROST score*: quality of the alignment of the query onto the candidate structure. A score larger than 9 means that the alignment is good, while a score less than 7 means the opposite. For values between 7 and 9, this score cannot be used to reach a decision.

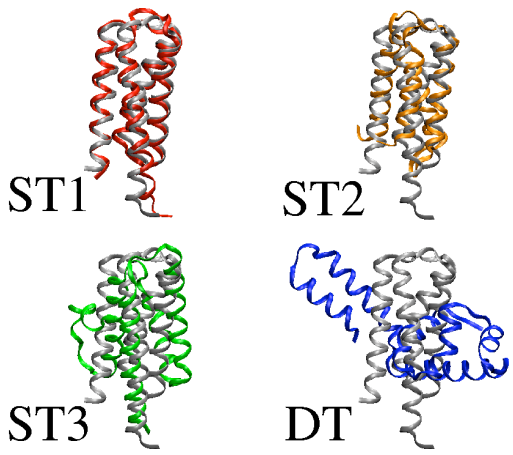

Figure 3: Superposition of the native structure *(grey)* with the **ST1** structure *(red.)*, the **ST2** structure *(orange)*, the **ST3** structure *(green)*, and the **DT** structure *(blue)*.

We took three cases for which the query sequence has been aligned with a structure similar to the native structure. For **ST1** and **ST2**, alignment is good or fair and the difference relies on the percentage of identity between the sequences. The **ST1** sequence is homolog with the query sequence, (% seq. Id. $> 20\%$), while we cannot be sure for **ST2**. **ST3** is certainly not an homolog

	NS/ST1	**NS/ST2**	**NS/ST3**	**NS/DT**
$\widehat{BF}$	1.34	1.22	2.42	2.76
$\widehat{\mathbb{P}}(\mathcal{M} = \mathbf{NS}\|\mathbf{x}^0)$	0.573	0.551	0.708	0.734

Table 3: Estimates of the Bayes factors between model **NS** and models **ST1**, **ST2**, **ST3**, and **DT**, and corresponding posterior probabilities of model **NS** based on an ABC-MC algorithm using $1.2\,10^6$ simulations and a tolerance ϵ equal to the 1% quantile of the distances.

of the query and the alignment is much poorer. For **DT**, the query sequence has been aligned with a structure that only share few structure elements with the native structure. It must be noted that alignments used to build **ST2**, **ST3**, and **DT** were scored in the FROST uncertainty zone, that is to say we cannot choose between these predictions based only on the results given by FROST as shown in Table 2. Differences between the native structure and the predicted structures are illustrated on Figure 3.

Using ABC-MC, we then estimate the Bayes factors between the model **NS** corresponding to the true structure and the models **ST1**, **ST2**, **ST3**, and **DT**, corresponding to the predicted structures. Estimated values and posterior probabilities of model **NS** are given for each case in Table 3. All Bayes factors are estimated to be larger than 1 so this indicates that data are always in favour of the native structure, when compared with one of the four alternatives. Moreover, the value is larger when models are more different. Note that the posterior probability of model **NS** is larger when the alternative model is **ST2** than when the alternative model is **ST1**.

Acknowledgments

The authors' research is partly supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2005 project ANR-05-BLAN-0196-01 Misgepop and by a grant from Region Ile-de-France.

References

[1] Beaumont, M., W. Zhang, and D. Balding. 2002. Approximate Bayesian Computation in Population Genetics. *Genetics* 162: 2025–2035.

[2] Gelman, A. and X. Meng. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Science* 13: 163–185.

[3] Green, P. and S. Richardson. 2002. Hidden Markov models and disease mapping. *J. American Statist. Assoc.* 92: 1055–1070.

[4] Ibanez, M. and A. Simo. 2003. Parametric estimation in Markov random fields image modeling with imperfect observations. A comparative study. *Pattern Recognition Letters* 24: 2377–2389.

[5] Marin, A., J. Pothier, K. Zimmermann, and J. Gibrat. 2002. FROST: a filterbased fold recognition method. *Proteins* 49: 493–509.

[6] Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré. 2003. Markov chain Monte Carlo without likelihoods. *Proc. National Acad. Sci. USA* 100(26): 15324–15328.

[7] Moeller, J., A. Pettitt, R. Reeves, and K. Berthelsen. 2006. An efficient MCMC algorithm method for distributions with intractable normalising constant. *Biometrika* 93: 451–458.

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

[9] Sali, A. and T. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234: 779–815.

[10] Taly, J., A. Marin, and J. Gibrat. 2008. Can molecular dynamics simulations help in discriminating correct from erroneous protein 3D models? *BMC Bioinformatics* 9: 6.

[11] Zhang, Y. and J. Skolnick. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* 57: 702–710.

4 Appendix: Binomial expectations

First, given $X \sim \mathcal{B}(N,p)$, we have :

$$\begin{aligned}\mathbb{E}\left[\frac{N-X+1}{X+1}\right] &= \mathbb{E}\left[\frac{N+2}{X+1}\right] - \mathbb{E}\left[\frac{X+1}{X+1}\right] \\ &= (N+2)\mathbb{E}\left[\frac{1}{X+1}\right] - 1\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[\frac{1}{1+X}\right] &= \sum_{k=0}^{N} \frac{1}{1+k}\binom{N}{k} p^k (1-p)^{N-k} \\ &= \frac{(1-p)^{N+1}}{p} \sum_{k=0}^{N} \frac{1}{k+1}\binom{N}{k} \left\{\frac{p}{1-p}\right\}^{k+1} \\ &= \frac{(1-p)^{N+1}}{p} \sum_{k=0}^{N} \binom{N}{k} \int_0^{\frac{p}{1-p}} x^k \,\mathrm{d}x \\ &= \frac{(1-p)^{N+1}}{p} \int_0^{\frac{p}{1-p}} \sum_{k=0}^{N} \binom{N}{k} x^k \,\mathrm{d}x \\ &= \frac{(1-p)^{N+1}}{p} \int_0^{\frac{p}{1-p}} (1+x)^N \,\mathrm{d}x \\ &= \frac{(1-p)^{N+1}}{p} \left[\frac{(1+x)^{N+1}}{N+1}\right]_0^{\frac{p}{1-p}} \\ &= \frac{1}{N+1}\frac{1}{p}\left\{1-(1-p)^{N+1}\right\}.\end{aligned}$$

Then

$$\begin{aligned}\mathbb{E}\left[\frac{N+2}{X+1}\right] &= \frac{N+2}{N+1}\frac{1}{p}\left\{1-(1-p)^{N+1}\right\} \\ &\xrightarrow[N\to\infty]{} \frac{1}{p}\end{aligned}$$