ABC likelihood-free methods for model choice in Gibbs random fields

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

Abstract. Gibbs random fields (GRF) 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 two proteins.

Keywords: Approximate Bayesian Computation, model choice, Gibbs Random Fields, Bayes factor, protein folding

1 Introduction

1.1 Gibbs random fields

We consider a finite set of sites $\mathcal{S} = \{1, \cdots, n\}$. At each site $i \in \mathcal{S}$, we observe $x_i \in \mathcal{X}_i$ where $\mathcal{X}_i$ is a finite set of states. $\mathcal{X} = \prod_{i=1}^{n} \mathcal{X}_i$ is the set of the configurations, $\mathbf{x} = (x_1, \cdots, x_n)$ corresponding to one configuration. We also consider an undirected graph $\mathcal{G} = (E(\mathcal{G}), V(\mathcal{G}))$ on $\mathcal{S}$, $V(\mathcal{G})$ being a vertex set and $E(\mathcal{G})$ an edge set. The sites i and $i^{'}$ are said neighbours (denoted $i \sim i^{'}$) if $(i, i^{'}) \in E(\mathcal{G})$, in other words, if there is a vertex between i and $i^{'}$. A clique c is a subset of $\mathcal{S}$ where all elements are mutual neighbours (Darroch et al. 1980). We denote by $\mathcal{C}$ the set of all cliques of the undirected graph $\mathcal{G}$.

---

[*]INRA Jouy-en-Josas, MIG, CEREMADE, Université Paris Dauphine & CREST, INSEE, France, mailto:aude.grelaud@jouy.inra.fr

[†]CEREMADE, Université Paris Dauphine & CREST, INSEE, France, mailto:xian@ceremade.dauphine.fr

[‡]Institut de Mathématiques et Modélisation de Montpellier, Université Montpellier 2& CREST, INSEE, France, mailto:jean-michel.marin@univ-montp2.fr

[§]INRA Jouy-en-Josas, MIG, France, mailto:francois.rodolphe@jouy.inra.fr

[¶]INRA Jouy-en-Josas, MIG, France, mailto:jean-francois.taly@jouy.inra.fr

In the finite framework previously adopted, Gibbs Random Fields (GRFs) are probabilistic models associated with densities (with respect to the counting measure)

$$f(\mathbf{x}) = \frac{1}{Z}\exp\{-U(\mathbf{x})\} = \frac{1}{Z}\exp\left\{-\sum_{c\in\mathcal{C}}U_c(\mathbf{x})\right\}, \tag{1}$$

where $U(\mathbf{x}) = \sum_{c\in\mathcal{C}}U_c(\mathbf{x})$ is the potential and Z is the corresponding normalising constant

$$Z = \sum_{\mathbf{x}\in\mathcal{X}}\exp\left\{-\sum_{c\in\mathcal{C}}U_c(\mathbf{x})\right\}.$$

If the density f of a Markov Random Field (MRF) is everywhere positive, then the Hammersley-Clifford theorem establishes that there exists a GRF representation of this MRF (Besag 1974).

We consider here GRF with potential $U(\mathbf{x}) = -\boldsymbol{\theta}^{\mathrm{T}}S(\mathbf{x})$ where $\boldsymbol{\theta}\in\mathbb{R}^p$ is a scale parameter, $S(\cdot)$ is a function taking values in $\mathbb{R}^p$. $S(\mathbf{x})$ is defined on the cliques of the neighbourhood system in that $S(\mathbf{x}) = \sum_{c\in\mathcal{C}}S_c(\mathbf{x})$. In that case, we have

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{1}{Z_{\boldsymbol{\theta}}}\exp\{\boldsymbol{\theta}^{\mathrm{T}}S(\mathbf{x})\}, \tag{2}$$

the normalising constant $Z_{\boldsymbol{\theta}}$ now depends on the scale parameter $\boldsymbol{\theta}$.

1.2   Bayesian model choice

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. Numerical approximations bypassing this constant like path sampling (Gelman and Meng 1998), pseudo likelihood (Besag 1975) or those based on an auxiliary variable (Møller et al. 2006) are not always available either because they require heavy computations or because they are not accurate enough in the case of the pseudo-likelihood. In particular, selecting a model with sufficient statistic S_0 taking values in $\mathbb{R}^{p_0}$ versus a model with sufficient statistics S_1 taking values in $\mathbb{R}^{p_1}$ relies on the Bayes factor corresponding to the priors π_0 and π_1 on the respective parameter spaces

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

$$\int \exp\{\boldsymbol{\theta}_1^{\mathrm{T}}S_1(\mathbf{x})\}/Z_{\boldsymbol{\theta}_1,1}\pi_1(\mathrm{d}\boldsymbol{\theta}_1)$$

but this quantity is not easily computable. One faces the same computational difficulties with the posterior probabilities of the models since they also depend on those unknown

constants. To properly approximate those posterior quantities, we thus propose an alternative resolution based on likelihood-free techniques such as Approximate Bayesian Computation (ABC) (Pritchard et al. 1999) and we show how ABC is naturally tuned for this purpose by providing a direct estimator of the Bayes factor.

From a modelling perspective, GRFs are used to model the dependency within spatially correlated data, with applications in epidemiology (Green and Richardson 2002) and image analysis (Ibanez and Simo 2003), among others (Rue and Held 2005). They often use a Potts model defined by a sufficient statistic S taking values in $\mathbb{R}$ in that

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

where $\sum_{i' \sim i}$ indicates that the summation is taken over all the neighbour pairs. In that case, $\mathcal{X} = \{1, \cdots, K\}^n$, $K = 2$ corresponding to the Ising model, and θ is a scalar. $S(\cdot)$ therefore monitors the number of identical neighbours over $\mathcal{X}$.

1.3 Plan

For a fixed neighbourhood or model, 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. In section 2, we describe the main likelihood-free algorithms before proposing a procedure based on an ABC algorithm aimed at selecting a model. Then, we show how to improve the accuracy of this approximation using an importance sampling procedure. In section 3, we 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 Approximate Bayesian Computation

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 Pritchard et al. (1999) and expanded in Beaumont et al. (2002) and Marjoram et al. (2003)] is based can be briefly described as follows: given a dataset $\mathbf{x}^0 = (x_1, \cdots, x_n)$ 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 jointly a value θ^* from the prior, $\theta^* \sim \pi(\cdot)$, and a value $\mathbf{x}^*$ 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

Exact rejection algorithm:

1. Generate θ^* from the prior π.

2. Generate $\mathbf{x}^*$ from the model $f(\cdot|\theta^*)$.

3. Accept θ^* if $\mathbf{x}^* = \mathbf{x}^0$, otherwise, start again in 1.

This solution is not approximative in that the output is truly simulated from the posterior distribution $\pi(\theta|\mathbf{x}^0) \propto f(\mathbf{x}^0|\theta)\pi(\theta)$ since $(\theta^*, \mathbf{x}^*) \sim \pi(\theta^*)\mathbb{I}_{\{\mathbf{X}^*=\mathbf{X}^0\}}f(\mathbf{x}^*|\theta)$. 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 ρ, given a fixed tolerance level ϵ, $\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon$. The distance ρ is open to choice but is usually an Euclidean distance $\rho(\mathbf{x}^*, \mathbf{x}^0) = \sum_{i=1}^n (x_i^* - x_i^0)^2$ (see Beaumont et al. (2002) or Blum and François (2008)). The corresponding ϵ-tolerance rejection algorithm is then

ϵ-tolerance rejection algorithm:

1. Generate θ^* from the prior π.

2. Generate $\mathbf{x}^*$ from the model $f(\cdot|\theta^*)$.

3. Accept θ^* if $\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon$, otherwise, start again in 1.

This approach is obviously approximative when $\epsilon \neq 0$. The output from the ϵ-tolerance rejection algorithm is thus associated with the distribution

$$\pi(\theta|\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon) \propto \pi(\theta)\mathbb{P}_\theta(\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon)$$

with $\mathbb{P}_\theta(\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon) = \int \mathbb{I}_{\{\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon\}} f(\mathbf{x}^*|\theta^*)d\mathbf{x}^*$. The choice of ϵ is therefore paramount for good performances of the method. If ϵ is too large, the approximation is poor; when $\epsilon \to \infty$, it amounts to simulating from the prior since all simulations are accepted (as $\mathbb{P}_\theta(\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon) \to 1$ when $\epsilon \to \infty$). If ϵ is sufficiently small, $\pi(\theta|\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon)$ is a good approximation of $\pi(\theta|\mathbf{x}^0)$. There is no approximation when $\epsilon = 0$, since the ϵ-tolerance rejection algorithm corresponds to the exact rejection algorithm, but the acceptance probability may be too low to be practical. Selecting the "right" ϵ is thus crucial. It is customary to pick ϵ as an empirical quantile of $\rho(\mathbf{x}^*, \mathbf{x}^0)$ when $\mathbf{x}^*$ is simulated from the marginal distribution $\mathbf{x}^* \propto \int \pi(\theta)\mathbb{P}_\theta(\rho(\mathbf{x}^*, \mathbf{x}^0) < \epsilon)d\theta$, and the choice is often the corresponding 1% quantile (see, for instance Beaumont et al. (2002) or Blum and François (2008)). Wilkinson (2008) proposes to replace the approximation by an exact simulation based on a convolution with an arbitrary kernel.

The data $\mathbf{x}^0$ usually being of a large dimension, another level of approximation is enforced within the true ABC algorithm, by replacing the distance $\rho(\mathbf{x}^*, \mathbf{x}^0)$ with a

corresponding distance between summary statistics $\rho(S(\mathbf{x}^*), S(\mathbf{x}^0))$ (Beaumont et al. 2002). When S is a sufficient statistic, this step has no impact on the approximation since $\pi(\theta|\rho(S(\mathbf{x}^*), S(\mathbf{x}^0))) = \pi(\theta|\rho(\mathbf{x}^*, \mathbf{x}^0))$. In practice, it is rarely the case that a sufficient statistic of low dimension is available when implementing ABC (see Beaumont et al. (2002) or Blum and François (2008)). As it occurs, the setting of model choice among Gibbs random fields is an exception in that it allows for such a beneficial structure, as will be shown 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:

ABC algorithm:

1. Generate θ^* from the prior π.

2. Generate $\mathbf{x}^*$ from the model $f(\cdot|\theta^*)$.

3. Compute the distance $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.

4. Accept θ^* if $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$, otherwise, start again in 1.

2.2   Model choice via ABC

In a model choice perspective, we face M Gibbs random fields in competition, each model m being associated with sufficient statistic 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 space $\Theta = \cup_{m=0}^{M-1}\{m\} \times \Theta_m$ that includes the model index $\mathcal{M}$, 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 $S(\mathbf{x})$ is a sufficient statistic 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

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

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

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

is the cardinality 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 $S(\mathbf{x})$ 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.

Note that when we consider M models from generic exponential families, this property of the concatenated sufficient statistic rarely holds. For instance, if under model $\mathcal{M} = 0$, $x_i|\theta_0 \overset{iid}{\sim} \mathcal{P}(\theta_0)$ and under model $\mathcal{M} = 1$, $x_i|\theta_1 \overset{iid}{\sim} \mathcal{G}eo(\theta_1)$, this property is not satisfied since the distribution of $\mathbf{x}$ given the common $S(\mathbf{x}) = \sum_{i=1}^{n} x_i$ in the first model

$$
h_0(\mathbf{x}|S(\mathbf{x})) = \left[\sum_{\tilde{\mathbf{x}} \in \mathcal{X}: S(\tilde{\mathbf{x}})=s} \frac{1}{\prod_{i=1}^{n} \tilde{x}_i!}\right]^{-1} \frac{1}{\prod_{i=1}^{n} x_i!}
$$

is different from the distribution of $\mathbf{x}$ given $S(\mathbf{x})$ in the other one

$$
h_1(\mathbf{x}|S(\mathbf{x})) = \frac{1}{n(S(\mathbf{x}))}.
$$

As a consequence, $S(\mathbf{x})$ is not sufficient for the parameter $\mathcal{M}$.

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

ABC algorithm for model choice (ABC-MC):

1. Generate m^* from the prior $\pi(\mathcal{M} = m)$.

2. Generate $\theta_{m^*}^*$ from the prior $\pi_{m^*}(\cdot)$.

3. Generate $\mathbf{x}^*$ from the model $f_{m^*}(\cdot|\theta_{m^*}^*)$.

4. Compute the distance $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.

5. Accept $(\theta^*_{m^*}, m^*)$ if $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon$, otherwise, start again in 1.

Simulating a data set $\mathbf{x}^*$ from $f_{m^*}(\cdot|\theta^*_{m^*})$ at step 3 is often non-trivial for GRFs. For the special case of the Ising model considered in the examples below, there have been many developments from Besag (1974) to Møller and Waagepetersen (2003) that allow for exact simulation via perfect sampling. We refer the reader to Häggström (2002), Møller (2003) and Møller and Waagepetersen (2003), for details of this simulation technique and for a discussion of its limitations. For other GRFs it is often possible to use a Gibbs sampler updating one clique at a time conditional on the others. This solution was implemented for the biophysical example of Section 3.2.

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 $(\theta^{i*}_{m^{i*}}, m^{i*})$ $(1 \le i \le N)$ 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

$$BF_{m_0/m_1}(\mathbf{x}^0) = \frac{\mathbb{P}(\mathcal{M} = m_0|\mathbf{x}^0)}{\mathbb{P}(\mathcal{M} = m_1|\mathbf{x}^0)} \frac{\pi(\mathcal{M} = m_1)}{\pi(\mathcal{M} = m_0)} \tag{3}$$

$$= \frac{\int f_{m_0}(\mathbf{x}^0|\theta_0)\pi_0(\theta_0)\pi(\mathcal{M} = m_0)\mathrm{d}\theta_0}{\int f_{m_1}(\mathbf{x}^0|\theta_1)\pi_1(\theta_1)\pi(\mathcal{M} = m_1)\mathrm{d}\theta_1} \frac{\pi(\mathcal{M} = m_1)}{\pi(\mathcal{M} = m_0)}. \tag{4}$$

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 $\sharp\{m^{i*} = m_1\} \ne 0$. 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 $N_0 = \sharp\{m^{i*} = m_0\}$, $N_1 = \sharp\{m^{i*} = m_1\}$ then 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

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

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.

$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ can be seen as the ratio of the posterior means on the model probabilities p under a $\mathcal{D}ir(1, \cdots, 1)$ prior. In fact, if we denote $N_j = \sharp\{m^{i*} = m_j\}$, $N = \sum_{j=0}^{M-1}$ then the vector $(N_1, \cdots, N_M)$ has a multinomial distribution

$$(N_0, \cdots, N_{M-1} | p_0, \cdots, p_{M-1}) \sim \mathcal{M}(N; p_0, \cdots, p_{M-1}).$$

The corresponding posterior distribution on p is a $\mathcal{D}ir(1 + N_0, \cdots, 1 + N_{M-1})$ and

$$\widehat{BF}_{m_0/m_1}(\mathbf{x}^0) = \frac{\mathbb{E}[p_0 | N_0, \cdots, N_{M-1}]}{\mathbb{E}[p_1 | N_0, \cdots, N_{M-1}]} = \frac{N_0 + 1}{N_1 + 1}$$

is a consistent estimate of $BF_{m_0/m_1}(\mathbf{x}^0)$.

Since the distribution of the sample $(\theta_{m^{i*}}^{i*}, m^{i*})_{(1 \leq i \leq N)}$ is not exactly $\pi\{(\mathcal{M}, \theta_0, \ldots, \theta_{M-1}) | \mathbf{x}^0\}$ but $\pi\{(\mathcal{M}, \theta_0, \ldots, \theta_{M-1}) | \rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon\}$, the Bayes factor should be written as

$$
\begin{aligned}
BF_{m_0/m_1}(\mathbf{x}^0) &= \frac{\mathbb{P}(\mathcal{M} = m_0 | \rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon)}{\mathbb{P}(\mathcal{M} = m_1 | \rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon)} \frac{\pi(\mathcal{M} = m_1)}{\pi(\mathcal{M} = m_0)} \\
&= \frac{\int \pi\{(\mathcal{M} = m_0, \theta_0) | \rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon\} \, d\theta_0 \, \pi(\mathcal{M} = m_1)}{\int \pi\{(\mathcal{M} = m_1, \theta_1) | \rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon\} \, d\theta_1 \, \pi(\mathcal{M} = m_0)} \\
&= \frac{\int \mathbb{P}_{\theta_0}(\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon)\pi_0(\theta_0)\pi(\mathcal{M} = m_0)d\theta_0 \, \pi(\mathcal{M} = m_1)}{\int \mathbb{P}_{\theta_1}(\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon)\pi_1(\theta_1)\pi(\mathcal{M} = m_1)d\theta_1 \, \pi(\mathcal{M} = m_0)} \\
&= \frac{\int \left[\int f_{m_0}(\mathbf{x}^* | \theta_0)\pi_0(\theta_0)d\theta_0\right] \mathbb{I}_{\{\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon\}} d\mathbf{x}^*}{\int \left[\int f_{m_1}(\mathbf{x}^* | \theta_1)\pi_1(\theta_1)d\theta_1\right] \mathbb{I}_{\{\rho(S(\mathbf{x}^0), S(\mathbf{x}^*)) < \epsilon\}} d\mathbf{x}^*}
\end{aligned}
$$

When $\epsilon = 0$ and $S(\mathbf{x})$ is a sufficient statistic, this expression corresponds to equation (3).

2.3 Two step ABC

The above estimator $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ is rather unstable (i.e. it 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 $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ 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. 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. Note, however, that, when there are more than two models, using these approximations to perform Bayesian model averaging can be dangerous.

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 GRF, the first one with a trivial neighbourhood structure and the other one with a nearest neighbourhood structure. 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 binary variables. Under model $\mathcal{M} = 0$, the GRF representation of the Bernoulli distribution $\mathcal{B}(\exp(\theta_0)/\{1 + \exp(\theta_0)\})$ 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 posterior probability of this model is available since, when $S_0(\mathbf{x}) = s_0$ $(s_0 \neq 0)$, the marginal probability 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 GRF 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

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

with $S_1(\mathbf{x}) = \sum_{i=2}^{n} \mathbb{I}_{\{x_i = x_{i-1}\}}$ being the sufficient statistic and $Z_{\theta_1,1} = 2(1 + e^{\theta_0})^{n-1}$ being the normalising constant in that case. Under a uniform prior $\theta_1 \sim \mathcal{U}(0,6)$, the posterior probability of this model is once again available, the likelihood being of the same form as when $\mathcal{M} = 0$. The bounds of the prior distributions on θ_0 and θ_1 were chosen to avoid data sets consisting in a sequence of n identical values since it is impossible to distinguish model 0 and model 1 in that case.

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 $\mathbf{x}^0 = (x_1, \cdots, x_n)$ with $n = 100$ under each model, using parameters simulated from the priors and computed the exact posterior probabilities and the Bayes factors in both cases. For each of those 2000 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 exactly simulated from the joint distribution and a random number of those were accepted when $S(\mathbf{x}^*) = S(\mathbf{x}^0)$. (In the worst case scenario, the number of acceptances was 12!) As shown on the left graph of Figure 1, the fit of the approximate posterior probabilities is 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 $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$, ρ being the Euclidean distance, the results are similar when $\mathbb{P}(\mathcal{M} = 0|\mathbf{x}^0)$ is close to 0, 1 or 0.5, and we observe a slight difference for other 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 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. If we classify the values of $BF_{m_0/m_1}(\mathbf{x}^0)$ and $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ according to the Jeffreys' scale, we observe that the Bayes factor and its approximation belong to the same category (1903 simulated data sets are on the diagonal of Table 1) or to very close categories. Once more, using a tolerance ϵ equal to the 1% quantile does not bring much difference in the output, Table 2 shows that the Bayes factor and its estimation still belong to the same category for 1805 simulated data sets. The approximative Bayes factor is slightly less discriminative in that case, since the slope of the cloud is less than the unitary slope of the diagonal on the right graph of Figure 2; $BF_{m_0/m_1}(\mathbf{x}^0)$ and $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ lead to the selection of the same model, but with a lower degree of confidence for the second one (Table 2). The boxplots in Figure
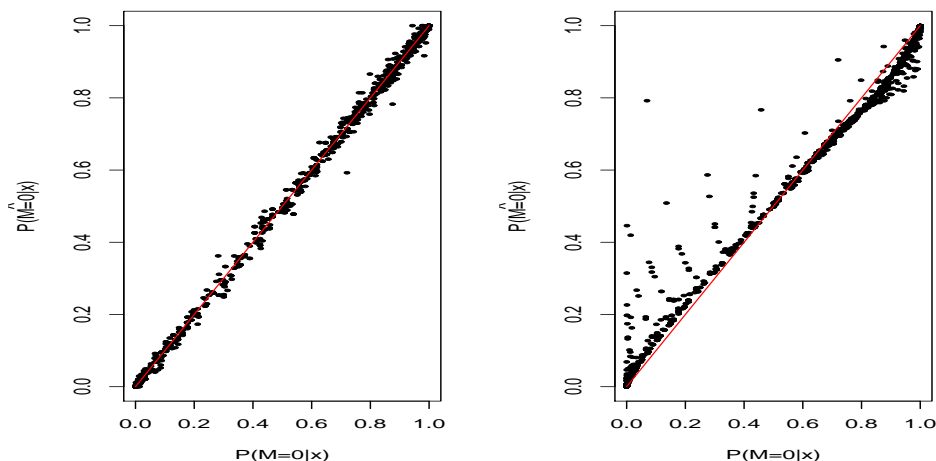
Figure 1: *(left)* Comparison of the true $\mathbb{P}(\mathcal{M} = 0|\mathbf{x}^0)$ with $\widehat{\mathbb{P}}(\mathcal{M} = 0|\mathbf{x}^0)$ over $2,000$ simulated sequences and 4×10^6 proposals from the prior. The solid/red line is the diagonal. *(right)* Same comparison when using a tolerance ϵ corresponding to the 1% quantile on the distances.

3 compare the distributions of the ratios $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)/BF_{m_0/m_1}(\mathbf{x}^0)$ in the exact case and using a tolerance equal to the 1% quantile on the distances. As reported in Table 3, the median is very close to 1 in both cases. The ratio takes more often extreme values in the exact case. Once more, this is a consequence of the poor estimation of the Bayes factor when the acceptance rate is small. 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

The numerous genome sequences now available provide a huge amount of protein sequences whose functions remain unknown. A classical strategy is to determine the tridimensional (3D) structure of a protein, also called *fold*, as it provides important and valuable information about its function. Experimental methods, like those based on X-ray diffraction or nuclear magnetic resonance, provide accurate descriptions of 3D-structures, but are time consuming. As an alternative, computational methods have been proposed to predict 3D structures.

These latter methods mostly rely on homology (two proteins are said to be homologous if they share a common ancestor). In fact, homologous proteins often share similar function and, as function is controlled by structure, similar structure. When the protein under study, hereafter called the query protein, can be considered as homologous with
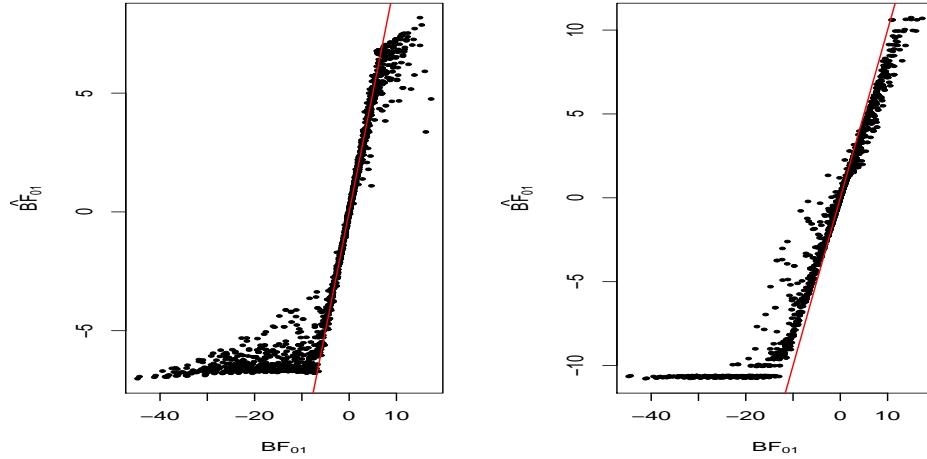
Figure 2:   *(left)* Comparison of the true $BF_{m_0/m_1}(\mathbf{x}^0)$ with $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ (in logarithmic scales) over $2,000$ simulated sequences and 4×10^6 proposals from the prior. The solid/red line is the diagonal.   *(right)* Same comparison when using a tolerance corresponding to the 1% quantile on the distances.
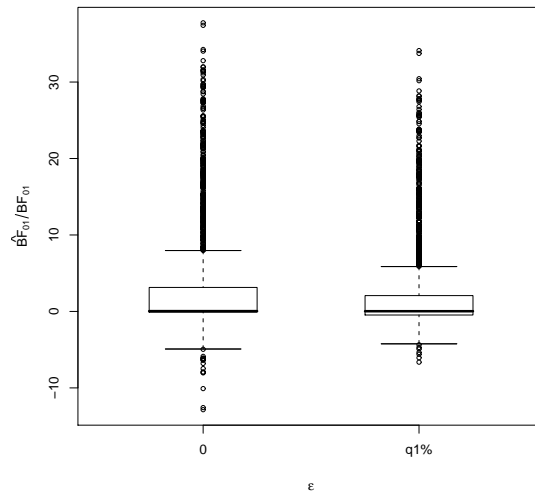

Figure 3:   *(left)* Boxplots of the ratios $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)/BF_{m_0/m_1}(\mathbf{x}^0)$ (in logarithmic scales) in the exact case and using a tolerance equal to the 1% quantile on the distances over $2,000$ simulated sequences and 4×10^6 proposals from the prior.

	$m=1$ dec.	$m=1$ str.	$m=1$ sub.	$m=1$ weak	$m=0$ weak	$m=0$ sub.	$m=0$ str.	$m=0$ dec.
$m=1$, dec.	778	9	0	0	0	0	0	0
$m=1$, str.	2	79	0	0	0	0	0	0
$m=1$, sub.	0	7	53	0	0	0	0	0
$m=1$, weak	0	0	2	63	0	7	0	0
$m=0$, weak	0	0	0	22	103	7	0	0
$m=0$, sub.	0	0	0	0	1	103	23	0
$m=0$, str.	0	0	0	0	0	5	177	6
$m=0$, dec.	0	0	0	0	0	0	13	547

Table 1: Comparison of the decisions based on $BF_{m_0/m_1}(\mathbf{x}^0)$ and on $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ using $\epsilon = 0$ according to the Jeffreys' scale (dec.: decisive $\log(BF_{m_0/m_1}(\mathbf{x}^0)) > 2$, str.: strong $1 < \log(BF_{m_0/m_1}(\mathbf{x}^0)) < 2$, sub.: substantial $0.5 < \log(BF_{m_0/m_1}(\mathbf{x}^0)) < 1$, weak $0 < \log(BF_{m_0/m_1}(\mathbf{x}^0)) < 0.5$).

	$m=1$ dec.	$m=1$ str.	$m=1$ sub.	$m=1$ weak	$m=0$ weak	$m=0$ sub.	$m=0$ str.	$m=0$ dec.
$m=1$, dec.	740	39	5	2	0	0	1	0
$m=1$, str.	0	64	14	2	1	0	0	0
$m=1$, sub.	0	0	39	19	2	0	0	0
$m=1$, weak	0	0	0	61	3	0	1	0
$m=0$, weak	0	0	0	2	128	2	0	0
$m=0$, sub.	0	0	0	0	2	123	1	1
$m=0$, str.	0	0	0	0	0	26	161	1
$m=0$, dec.	0	0	0	0	0	0	71	489

Table 2: Comparison of the decisions based on $BF_{m_0/m_1}(\mathbf{x}^0)$ and on $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)$ according to the Jeffreys' scale, using a tolerance ϵ corresponding to the 1% quantile of the distances (dec.: decisive $\log(BF_{m_0/m_1}(\mathbf{x}^0)) > 2$, str.: strong $1 < \log(BF_{m_0/m_1}(\mathbf{x}^0)) < 2$, sub.: substantial $0.5 < \log(BF_{m_0/m_1}(\mathbf{x}^0)) < 1$, weak $0 < \log(BF_{m_0/m_1}(\mathbf{x}^0)) < 0.5$).

	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$
$\epsilon = 0$	0.914	1.041	22.9
$\epsilon = q_{1\%}$	0.626	1.029	7.9

Table 3: Quantiles of the ratios $\widehat{BF}_{m_0/m_1}(\mathbf{x}^0)/BF_{m_0/m_1}(\mathbf{x}^0)$ in the exact case and using a tolerance ϵ equal to the 1% quantile of the distances.

Table 4: Classification of amino acids into hydrophilic and hydrophobic groups.

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

another protein, a prediction of its 3D structure based on the structure of its homolog can be built.

First, one compares the sequence of the query protein with a data bank of sequences of proteins of known structures but sequence similarity is often too low to assess homology with sufficient certainty. Because of selection pressure on the function, structures are more conserved over time than sequences. *Threading* methods consist in aligning the query sequence onto a set of structures representative of all known folds. The sequence of the query is threaded onto the candidate structures in order to find the most compatible one. A score (a fitting criterion) is computed for each proposal. Structures displaying sufficiently high scores, if any, are chosen as the corresponding protein can be said homologous with the query protein.

It may happen that both information based on sequence similarity and threading score are not sufficient to access protein homology and consequently, to select a 3D structure. Our aim is to use extra information to help making a decision, if necessary. We use here the fact that amino acids in close contact in the 3D structure often share similar (or complementary) biochemical properties. In the example we discuss in this section, we use hydrophobicity as a clustering factor since hydrophobic amino-acids are mostly buried inside the 3D structure, and hydrophilic ones exposed to water. This effect is observed in almost all proteins.

From a formal perspective, each structure can be represented by a graph where a node represents one amino-acid of the protein and an edge between two nodes indicates that both amino-acids are in close contact in the folded protein (hence are neighbours). Labels are allocated to each node, associated with hydrophobicity of amino-acids (amino-acids are classified as hydrophobic or hydrophilic according to Table 4). Then, a Gibbs random field, more precisely an Ising model, can be defined on each graph. When several structures are proposed by a threading method, the ABC-MC algorithm is then available to select the most likely structure.

We applied this procedure to proteins of known structures (here called the native structures) **1tqgA**, involved into signal transduction processes in the bacterium *Thermotoga maritima* and **1k77A** which is a putative oxygenase from *Escherichia coli*. In these studies, the sequences were treated as queries, since our purpose was to evaluate if our idea could help in real situations.

We used FROST (Marin et al. 2002), a software dedicated to threading, and MODELLER (Sali and Blundell 1993) to find the candidate structures and KAKSI (Martin et al. 2005) to build the graphs. All candidate structures were picked up from the Pro-
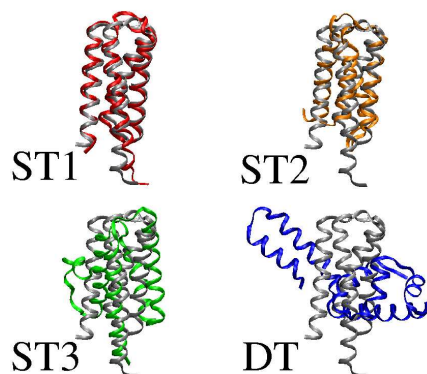
Figure 4: Superposition of the native structure of **1tqgA** *(grey)* with the **ST1** structure *(red)*, the **ST2** structure *(orange)*, the **ST3** structure *(green)*, and the **DT** structure *(blue)*.

tein Data Bank (http://www.rcsb.org/pdb/home/home.do). FROST provides the best alignment of the query sequence onto a structure, based on score optimisation, and the final score measures alignment quality. 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. Additionally, FROST calculates the percentages of identity between query and candidate sequences; sequences with a percentage of sequence identity higher than $> 20\%$ can be considered as homologous.

As the native structures were known, similarities between candidate and native structures could be assessed, here by the TM-score, (Zhang and Skolnick 2004). A score larger than 0.4 implies both structures are similar and a score less than 0.17 means that the prediction is nothing more than a random selection from the PDB library.

For each query protein, we selected four candidates, called **ST1**, **ST2**, **ST3** and **DT**, covering the whole spectrum of predictions that can be generated by protein threading, from good to very poor (Taly et al. 2008) as described in Table 5 and 6. We selected essentially candidate structures for which no decision could have been made since they were scored in the FROST uncertainty zone. According to the TM-score, **ST1** and **ST2** are considered as similar to the native structure, while **ST3** and **DT** are not. For **ST1** and **ST2**, the alignment of the query sequence onto the candidate structure is good or fair; sequence similarity is higher for **ST1** than **ST2**. **ST3** is a poorer candidate since it is certainly not an homolog of the query and the alignment is much poorer. For **DT**, the query sequence has been aligned with a structure that only shares few structural elements with the native structure. Differences between the native structures and the corresponding predicted structures are illustrated in Figure 4 for **1tqgA** and in Figure 5 for **1k77A**.

Using ABC-MC, we then estimate the Bayes factors between model **NS** corresponding to the true structure and models **ST1**, **ST2**, **ST3**, and **DT**, correspond-
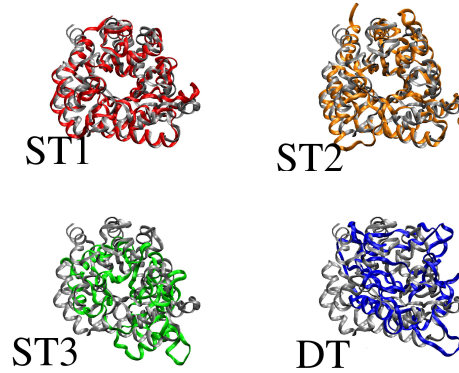
Figure 5: Superposition of the native structure of **1k77A** *(grey)* with the **ST1** structure *(red)*, the **ST2** structure *(orange)*, the **ST3** structure *(green)*, and the **DT** structure *(blue)*.

	% 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 5: Summary of the characteristics of our dataset for the protein **1tqgA**. *% seq. Id.*: percentage of identity with the query sequence. *TM-score*: similarity between a predicted structure and the native structure. *FROST score*: quality of the alignment of the query onto the candidate structure.

	% seq. Id.	TM-score	FROST score
1i60A (**ST1**)	16	0.69	8.9
1qtwA (**ST2**)	6	0.54	9.8
1qpoA1 (**ST3**)	9	0.29	9.3
1m4oA (**DT**)	7	0.17	8.3

Table 6: Summary of the characteristics of our dataset for the protein **1k77A**. *% seq. Id.*: percentage of identity with the query sequence. *TM-score*: similarity between a predicted structure and the native structure. *FROST score*: quality of the alignment of the query onto the candidate structure.

	NS/ST1	NS/ST2	NS/ST3	NS/DT
$\widehat{BF}$	1.34	1.22	2.42	2.76

Table 7: Estimates of the Bayes factors between model **NS** and models **ST1**, **ST2**, **ST3**, and **DT**, based on an ABC-MC algorithm using $1, 2 \times 10^6$ simulations and a tolerance ϵ equal to the 1% quantile of the distances for the query protein **1tqgA**.

	NS/ST1	NS/ST2	NS/ST3	NS/DT
$\widehat{BF}$	1.07	1.14	11997	14.24

Table 8: Estimates of the Bayes factors between model **NS** and models **ST1**, **ST2**, **ST3**, and **DT**, **NS** based on an ABC-MC algorithm using $1, 2 \times 10^6$ simulations and a tolerance ϵ equal to the 1% quantile of the distances for the query protein **1k77A**.

ing to the predicted structures. Each model is an Ising model with sufficient statistic $S_m(\mathbf{x}) = \sum_{i \sim i'}^{m} \mathbb{I}_{\{x_i = x_{i'}\}}$. The scalar parameter θ_m of the Ising model m is assumed to have a uniform prior on the interval $[0, 4]$. Simulated data sets were obtained by a standard Gibbs sampler. The Gibbs algorithm has been iterated 1000 times, which is a sufficient number of iterations for stabilisation. We picked ϵ as the empirical 1%-quantile of the Euclidean distance $\rho(S(\mathbf{x}^0), S(\mathbf{x}^*))$.

Estimated values for the Bayes factors of model **NS** against each alternative are given in Tables 7 and 8. As expected, all estimated Bayes factors are larger than 1 indicating that the data is always in favour of the native structure, when compared with one of the four alternatives and Bayes factors increase when the similarity between candidate and native structure is lower. Moreover, we can classify the candidate structures into two categories: for **ST1** and **ST2**, the evidence is weak in favour of the native structure while the evidence is substantial or strong when the alternative is **ST3** or **DT**. Thus our approach can distinguish similar from dissimilar structures, even when they were scored in the FROST uncertainty zone.

4   Discussion

This paper has hopefully demonstrated that the auxiliary variable technique that supports the ABC algorithm can be used to overcome the lack of closed-form normalising constants in Gibbs random field models and in particular in Ising models. The computation of Bayes factors can therefore follow from a standard Monte Carlo simulation that includes the model index without requiring advanced techniques like reversible jump moves (Robert and Casella 2004). The usual approximation inherent to ABC methods can furthermore be avoided due to the availability of a sufficient statistic across models. However, the toy example studied above shows that the accuracy of the approximation to the posterior probabilities and to the Bayes factor can be greatly improved by re-

sorting to the original ABC approach, since it allows for the inclusion of many more simulations. In the biophysical application to the choice of a folding structure for two proteins, we have also demonstrated that we can implement the ABC solution on realistic datasets and, in the examples processed there, that the Bayes factors allow for a ranking more standard methods do not.

Acknowledgements

References

Beaumont, M., W. Zhang, and D. Balding. 2002. Approximate Bayesian computation in population genetics. *Genetics* 162:2025–2035.   319, 320, 321

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statist. Society, Series B* 36:192–236.   318, 323

Besag, J. 1975. Statistical analysis of a non-lattice data. *The Statistician* 24(3):179–195.   318

Blum, M. G. B. and O. François. 2008. Highly tolerant likelihood-free Bayesian inference: An adaptive non-linear heteroscedastic model. *Statistics and Computing* (to appear).   320, 321

Darroch, J.N., S.L. Lauritzen, and T.P. Speed. 1980. Markov fields and log-linear interaction model for contingency tables. *Annals of Statistics* 8(3):522–539.   317

Gelman, A. and X. Meng. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Science* 13:163–185.   318

Green, P. and S. Richardson. 2002. Hidden Markov models and disease mapping. *J. American Statist. Assoc.* 92:1055–1070.   319

Häggström, O. (2002). Finite Markov Chains and Algorithmic Applications, volume 52, *Student Texts.* London Mathematical Society.   323

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.   319

Marin, A., J. Pothier, K. Zimmermann, and J. Gibrat. 2002. FROST: a filterbased fold recognition method. *Proteins* 49:493–509.   330

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. 319

Martin, J., G. Letellier, A. Marin, J.-F. Taly, A. de Brevern, and J.-F. Gibrat. 2005. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct. Biol.* 5:17. 330

Møller, J. 2003. *Spatial Statistics and Computational Methods*, volume 173 of *Lecture Notes in Statistics*. Springer-Verlag, New York. 323

Møller, J., A. Pettitt, R. Reeves, and K. Berthelsen. 2006. An efficient MCMC algorithm method for distributions with intractable normalising constant. *Biometrika* 93:451–458. 318

Møller, J. and R. Waagepetersen. 2003. *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton, FL. 323

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. 319

Robert, C.P. and G. Casella. 2004. *Monte Carlo Statistical Methods.* Springer-Verlag, New York. 333

Rue, H. and L. Held. 2005. *Gaussian Random Fields: Theory and Applications.* Chapman and Hall/CRC, Boca Raton, FL. 319

Sali, A. and T. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815. 330

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. 331

Toni, T., D. Welch, N. Strelkowa, A. Ipsen and M.P. Stumpf. 2008. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. Royal Society Interface* 6:187–202.

Wilkinson, R. D. 2008. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. arXiv:0811.3355 320

Zhang, Y. and J. Skolnick. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins* 57:702–710. 331