Bayesian regularization of the length of memory in reversible sequences

Sergio Bacallado

Department of Statistics, Stanford University, Stanford, CA, United States.

Vijay Pande

Department of Chemistry, Stanford University, Stanford, CA, United States.

Stefano Favaro

Department of Economics and Statistics, University of Torino and Collegio Carlo Alberto, Torino, Italy.

Lorenzo Trippa

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, United States.

Summary. Variable-order Markov chains have been used to model discrete sequential data in a variety of fields. A host of methods exist to estimate the history-dependent lengths of memory which characterize these models and to predict new sequences. In several applications, the data-generating mechanism is known to be reversible, but combining this information with the procedures mentioned is far from trivial. We introduce a Bayesian analysis for reversible dynamics, which takes into account uncertainty in the lengths of memory. The proposed model is applied to the analysis of molecular dynamics simulations and compared to several popular algorithms.

Keywords: Bayesian analysis, reinforced random walk, reversibility, variable-order Markov model.

1. Introduction

Time reversibility characterizes numerous stochastic models, from queueing networks (Kelly, 1979) to models of physical systems governed by reversible mechanics

(Van Kampen, 1992). This property is commonly associated to Markov chains. However, in several applications the Markovian assumption is restrictive, and one needs to model higher-order dependencies. Higher-order Markov models require the investigator to deal with the problem of a rapidly increasing number of unknown parameters. Here, we propose a Bayesian approach to this problem, tailored to reversible processes.

Our motivating application is the analysis of molecular dynamics simulations. These are computer experiments which mimic the structural transitions of macromolecules in time using a physics-based Hamiltonian operator. The simulation is a reversible Markov chain representing the trajectory of hundreds of atoms in a three dimensional space. This high-dimensional process is typically projected onto a partition of the space of molecular structures, to produce a discrete time series which is still reversible, even though it is not necessarily a Markov chain. An accurate characterization of the projected dynamics provides the biologist with estimates of the rates of transitions between biologically relevant states. Such estimates, with associated uncertainty measures, are also useful in the design of adaptive simulations which can significantly reduce the computational burden of these experiments (Prinz et al., 2011).

In variable-order Markov models, the length of memory is a function of the previously observed states. If all possible sequences of states are arranged in a context tree, where every branch is truncated, then it is only necessary to specify transition probabilities at the nodes where truncations occur. The length of each branch is a context-specific length of memory. The literature on variable-order Markov chains can be traced back to Rissanen (1983) and Weinberger et al. (1995) who developed algorithms based on context tree pruning. Statistical properties of these algorithms were developed by Bühlmann and Wyner (1999) and Bühlmann (2000). Begleiter et al. (2004) review relevant algorithms, such as Context Tree Weighting which is similar in spirit to Bayesian model averaging methods.

Bayesian nonparametric priors have been recently proposed for higher-order Markov chains. In the hierarchical Dirichlet language model (MacKay and Peto, 1995), there are transition distributions G_u out of every point u in $\cup_{i=1}^{n} \mathcal{X}^i$, where

$\mathcal{X}$ is the state space and n is the length of memory. If, for instance, $\mathcal{X}$ corresponds to the alphabet, then the prior for G_{word} is a Dirichlet distribution with mean identical to the ancestor G_{ord}. Teh (2006), Wood et al. (2009) and Mochihashi and Sumita (2008) developed successful nonparametric extensions of this construction for language modeling.

However, it would be difficult to combine the hierachical structure of these Bayesian models with our goal of doing inference for a reversible process. Annis et al. (2010) have shown that when the data-generating process is known to be reversible, models that enforce reversibility can have superior asymptotic properties. In order to develop a procedure that incorporates reversibility, we introduce the *random walk with amnesia*. This process generalizes the variable-order Markov model and is related to *Probabilistic Suffix Automata*, a construction introduced by Ron et al. (1996). Relying on the conjugate prior for reversible variable-order Markov chains introduced by Bacallado (2011), we define a Bayesian analysis for the random walk with amnesia. Our main contribution is an efficient procedure for Bayesian inference of reversible dynamics when the context-specific lengths of memory are unknown.

We note that an alternative, and widely used, modeling approach for sequential data is offered by hidden Markov models and extensions such as tiered hidden Markov models; see Cappé et al. (2005) for a comprehensive treatment. These models introduce memory through a latent process instead of explicitly modeling the dependence on history, like the methods considered in this paper. Reversible versions have been applied recently (Palla et al., 2014).

2. Reversible dynamics

2.1. The random walk with amnesia

We define a random walk with finite memory taking values in a finite space $\mathcal{X}$. Our random walker remembers only the last portion of his trajectory, and the length of this vector of states changes in time. At every step, the walker can either (i) lose the first element of his memory, or (ii) proceed to a new state in $\mathcal{X}$. For example,

when $\mathcal{X} = \{a, b, c\}$ a path could be

$$abc \overset{(i)}{\to} bc \overset{(i)}{\to} c \overset{(ii)}{\to} ca \overset{(ii)}{\to} cab, \tag{1}$$

where $\overset{(i)}{\to}$ denotes a transition of the first kind and $\overset{(ii)}{\to}$ a transition of the second kind.

We use an nth-order Markov chain $(X_i)_{i\in\mathbb{N}}$ on the enriched space $\mathcal{X}_s = \mathcal{X} \cup \{s_k; k = 1, \ldots, n\}$ to represent this process. The maximum number of $\mathcal{X}$-valued states that the random walker can remember is $n-1$. If for some $k > 0$, $x_1 = \cdots = x_k = s_k$ and $(x_{k+1}, \ldots, x_n) \in \mathcal{X}^{n-k}$ we say that $(x_1, \ldots, x_n)$ belongs to $\mathcal{Z} \subset \mathcal{X}_s^n$. This n-gram represents a state in which the random walker only remembers the sequence $(x_{k+1}, x_{k+2}, \ldots, x_n) \in \mathcal{X}^{n-k}$; the first k states $x_1, \ldots, x_k$ are equal to s_k, denoting a loss of memory. The sequence of n-grams belonging to $\mathcal{Z}$ in the process $(X_i)_{i\in\mathbb{N}}$ will mirror a trajectory like the one in (1). Definition 1 puts restrictions on the nth-order Markov chain $(X_i)_{i\in\mathbb{N}}$ which ensure that it represents the process described informally above. In the nth-order Markov chain $(X_i)_{i\in\mathbb{N}}$ the only random transitions occur from n-grams $(X_j, \ldots, X_{j+n-1})$ in $\mathcal{Z}$, while the remaining ones are deterministic. The random transitions determine whether (i) the random walker forgets the first state in her memory or (ii) adds a new $\mathcal{X}$ state to his trajectory. After each random transition, a series of deterministic transitions will brings the process to the subsequent n-gram in $\mathcal{Z}$.

Definition 1 *A random walk with amnesia is an nth-order Markov chain $(X_i)_{i\in\mathbb{N}}$ on the space $\mathcal{X}_s$. With probability 1, $(X_1, \ldots, X_n) \in \mathcal{Z}$. Given $(X_j, \ldots, X_{j+n-1}) = (x_1, \ldots, x_n) = x$, for any $j > 0$, the next state X_{j+n} satisfies the following constraints. If $x \in \mathcal{Z}$ and $x_1 = \cdots = x_k = s_k$, then X_{j+n} can be (i) s_{k+1}, provided $k < n$, or (ii) a state in $\mathcal{X}$, provided $k > 1$. If $x \notin \mathcal{Z}$, the state X_{j+n} is chosen deterministically according to the following rules:*

- *(a) if $x_1 = \cdots = x_{k-1} = s_k$ and $(x_k, \ldots, x_n) \in \mathcal{X}^{n-k+1}$ for some $1 < k \leq n$, then $X_{j+n} = s_{k-1}$;*
- *(b) if $x_{n-m} = \cdots = x_n = s_k$ and $x_{n-m-1} \neq s_k$ for some $m \leq k-2 \leq n-2$, then $X_{j+n} = s_k$;*

(c) if $x_{m+1} = \cdots = x_{m+k} = s_k$ *and* $x_m \neq s_k$ *for some* $0 < k < k + m \leq n$*, then* $X_{j+n} = x_1$.

If we consider the path in (1) and set $n = 4$, then, by Definition 1, the corresponding transitions of the nth-order Markov chain $(X_i)_{i\in\mathbb{N}}$ on the enriched space $\mathcal{X}_s$ are

$$\mathbf{s_1abc} \overset{(i)}{\to} abcs_2 \to bcs_2s_2 \to cs_2s_2b \to \mathbf{s_2s_2bc} \overset{(i)}{\to} s_2bcs_3 \to bcs_3s_3 \to cs_3s_3s_3 \to \mathbf{s_3s_3s_3c} \overset{(ii)}{\to}$$
$$s_3s_3ca \to s_3cas_2 \to cas_2s_2 \to as_2s_2c \to \mathbf{s_2s_2ca} \overset{(ii)}{\to} s_2cab \to cabs_1 \to abs_1c \to bs_1ca \to \mathbf{s_1cab},$$

where $\overset{(i)}{\to}$ and $\overset{(ii)}{\to}$ denote random transitions from n-grams in $\mathcal{Z}$, displayed in bold, and $\to$ denotes deterministic transitions. In this example $(X_1, X_2, \ldots) = (s_1, a, b, c, s_2, s_2, b, c, s_3, s_3, \ldots)$.

One can easily verify that the process $(X_i)_{i\in\mathbb{N}}$ only goes through n-grams covered by Definition 1. The law of the process is specified by the distribution of $(X_1, \ldots, X_n) \in \mathcal{Z}$ and the transition probabilities out of the n-grams in $\mathcal{Z}$.

Clearly, the definition of $(X_i)_{i\in\mathbb{N}}$ does not allow every possible path. However, if $(x_1, \ldots, x_r) \in \mathcal{X}_s^r$, $r > n$, is a realization of the process with $(x_r, \ldots, x_{r-n+1}) \in \mathcal{Z}$, then its inverse $(x_r, x_{r-1}, \ldots, x_1)$ is also an allowed path. In other words, if $(x_1, \ldots, x_r)$ is consistent with the deterministic constraints (a,b,c) in Definition 1 and $(x_r, \ldots, x_{r-n+1}) \in \mathcal{Z}$, then $(x_r, x_{r-1}, \ldots, x_1)$ is also consistent with these three requirements. For instance, with $n = 3$ and $\mathcal{X} = \{a, b, c\}$, the path $(x_1, \ldots, x_r) = (s_1, a, b, s_2, s_2, b, c, s_1)$ corresponds to a random walker in b who forgets a previous visit to a and then goes to c, while the reverse sequence $(s_1, c, b, s_2, s_2, b, a, s_1)$ corresponds to a random walker in b who forgets a previous visit to c and then goes to a. Every loss of memory in the forward path $(x_1, \ldots, x_r)$ is associated with a specular transition in $(x_r, \ldots, x_1)$ to an $\mathcal{X}$-valued state and viceversa. This will be a key property to introduce reversible random walks with amnesia in the sequel. It will also be useful in Section 3 to describe our Bayesian approach to infer the transition probabilties of $(X_i)_{i\in\mathbb{N}}$, under the assumption of reversibility, as a model-based reinforcement learning procedure.

An irreducible nth-order Markov chain $(X_i)_{i\in\mathbb{N}}$ on $\mathcal{X}_s$ is canonically represented by a *balanced* function $w_{n+1} : \mathcal{X}_s^{n+1} \to [0, \infty)$, which satisfies for any $(x_1, \ldots,$

$x_n) \in \mathcal{X}_s$,

$$\sum_{v \in \mathcal{X}_s} w_{n+1}(x_1, \ldots, x_n, v) = \sum_{v \in \mathcal{X}_s} w_{n+1}(v, x_1, \ldots, x_n) = w_n(x_1, \ldots, x_n) \quad \text{and} \tag{2}$$

$$\text{pr}(X_{n+m+1} | X_1, \ldots, X_{n+m}) = \frac{w_{n+1}(X_{1+m}, \ldots, X_{n+m+1})}{w_n(X_{m+1}, \ldots, X_{m+n})}. \tag{3}$$

Note that w_n is a stationary measure of this nth-order Markov chain. Without loss of generality, in what follows, we will assume $\sum_x w_{n+1}(x) = 1$.

The nth-order Markov chain $(X_i)_{i \in \mathbb{N}}$ is reversible if and only if

$$w_{n+1}(x_1, \ldots, x_{n+1}) = w_{n+1}(x_{n+1}, \ldots, x_1) \text{ for every } (x_1, \ldots, x_{n+1}) \in \mathcal{X}_s^{n+1}. \tag{4}$$

The following is a simple approach to define a function w_{n+1} consistent with this notion of reversibility. The resulting Markov chain is not necessarily a random walk with amnesia. Consider a cyclic sequence $(x_1, \ldots, x_r) \in \mathcal{X}_s^r$, with $r > n$ and $(x_1, \ldots, x_n) = (x_{r-n+1}, \ldots, x_r)$, and define

$$w_{n+1}(y_1, \ldots, y_{n+1}) = \frac{1}{2(r-n)} \times$$
$$\sum_{i=1}^{r-n} \mathbf{1}\left((x_i, \ldots, x_{i+n}) = (y_1, \ldots, y_{n+1})\right) + \mathbf{1}\left((x_i, \ldots, x_{i+n}) = (y_{n+1}, y_n, \ldots, y_1)\right)$$

for every n-gram $(y_1, \ldots, y_{n+1}) \in \mathcal{X}_s^{n+1}$, where $\mathbf{1}(\cdot)$ denotes an indicator function. The resulting function w_{n+1} satisfies equation (2) and allows one to specify the transition probabilities of a reversible nth-order Makov chain through the identity (3). This construction can be used to prove the following result.

Proposition 1 *There exist random walks with amnesia with a reversible parameter* w_{n+1}.

Proof. The random walk with amnesia is a class of nth-order Markov chains on $\mathcal{X}_s$ where certain transitions are not allowed. Therefore, it is sufficient to follow the constructive approach that we described in the previous paragraph to specify a reversible w_{n+1}. In this case we define w_{n+1} through a cyclic sequence $(x_1, \ldots, x_r)$, with $(x_1, \ldots, x_n) = (x_{r-n+1}, \ldots, x_r)$, which is in addition consistent with the constraints in Definition 1. The remark that, if $(x_1, \ldots, x_r)$ is consistent with the constraints in Definition 1, then so is the reversed sequence $(x_r, \ldots, x_1)$, completes the proof.

We note that if $w_{n+1} = \lambda \times w'_{n+1} + (1-\lambda) \times w''_{n+1}$, $\lambda \in (0,1)$, where both w'_{n+1} and w''_{n+1} are reversible parameters for random walks with amnesia, then w_{n+1} satisfies equations (2) and (4). This fact implies that reversible parameterizations for random walks with amnesia w_{n+1} constitute a convex space.

2.2. The observable random sequence

We use random walks with amnesia to model sequences with higher-order dependencies. Recall that the state space of the amnesia model is $\mathcal{X}_s$. We assume to observe only the sequence $(Z_j)_{j\in\mathbb{N}}$ of $\mathcal{X}$-valued states visited by our random walker. To make the definition precise, let $\tau_0 = n$ and $\tau_j = \min\{\tau; \tau > \tau_{j-1}, (X_{\tau-n}, ..., X_{\tau-1}) \in \mathcal{Z}, X_\tau \in \mathcal{X}\}$. The observable process is $Z_j := X_{\tau_j}$, $j \geq 1$. In the example displayed in (1) $Z_1 = a$ and $Z_2 = b$. The observable process does not need to be a Markov chain of any finite order, even though, as we show in the sequel, reversible variable order Markov chains are a special case.

Proposition 2 *Suppose that, given $X_1, \ldots, X_{n+1}$, the process $(X_i)_{i\in\mathbb{N}}$ is a reversible and irreducible random walk with amnesia with parameter w_{n+1}, and let*

$$pr(X_1 = x_1, \ldots, X_{n+1} = x_{n+1}) \propto w_{n+1}(x_1, \ldots, x_{n+1}) \mathbf{1}\left((x_1, \ldots, x_n) \in \mathcal{Z}, x_{n+1} \in \mathcal{X}\right). \tag{5}$$

Then the observable process $(Z_i)_{i\in\mathbb{N}}$ is stationary and reversible, i.e.

$$pr(Z_1 = z_1, \ldots, Z_m = z_m) = pr(Z_1 = z_m, \ldots, Z_m = z_1), \tag{6}$$

for any $m > 1$ and $z = (z_1, \ldots, z_m) \in \mathcal{X}^m$.

In the following, we neglect the condition in (5), as we are interested in estimating the transition probabilities in the latent process $(X_i)_{i\in\mathbb{N}}$ from observations which are not stationary. Nonetheless, assumption (5) is unnecessary to verify the following notion of reversibility.

Proposition 3 *Let $(X_i)_{i\in\mathbb{N}}$ be a reversible and irreducible random walk with amnesia. For any $m>1$ and $z = (z_1,\ldots,z_m) \in \mathcal{X}^m$, with probability 1,*

$$\lim_{k\to\infty} \sum_{i=1}^{k} \frac{\mathbf{1}(Z_i = z_1, \ldots, Z_{i+m-1} = z_m)}{k} = \lim_{k\to\infty} \sum_{i=1}^{k} \frac{\mathbf{1}(Z_i = z_m, \ldots, Z_{i+m-1} = z_1)}{k}. \tag{7}$$

The proofs of Propositions 2 and 3 are in the Supplementary Materials. We conclude this section showing that reversible variable order Markov chains are included within the family of observable processes $(Z_i)_{i\in\mathbb{N}}$ defined above. A variable order Markov chain $(U_i \in \mathcal{X})_{i\in\mathbb{N}}$ with maximum order $n-1$ and *histories* $\mathcal{H} \subset \cup_{i=1}^{n-1}\mathcal{X}^i$, satisfies the equality

$$\text{pr}(U_j = x_j | U_1 = x_1, \dots, U_{j-1} = x_{j-1}) = \text{pr}(U_j = x_j | U_{j-k} = x_{j-k}, \dots, U_{j-1} = x_{j-1})$$

for every $(x_1, \dots, x_j)$ whenever $(x_{j-k}, \dots, x_{j-1}) \in \mathcal{H}$. We will assume without loss of generality that $(U_1, \dots, U_{n-1})$ are fixed. We also assume that $(U_i)_{i\in\mathbb{N}}$ is reversible.

Proposition 4 *The sequence $(U_i)_{i\geq n}$ is identical in distribution to a sequence $(Z_i)_{i\in\mathbb{N}}$ of $\mathcal{X}$ valued states from a random walk with amnesia whose parameter $w_{n+1} : \mathcal{X}_s^{n+1} \to [0,\infty)$ is equal to zero in a subset $\mathcal{S}_\mathcal{H}$ of $\mathcal{X}_s^{n+1}$ specified by $\mathcal{H}$. A sequence $(x_1, \dots, x_{n+1})$ with $x_{n+1} \in \mathcal{X}$ belongs to $\mathcal{S}_\mathcal{H}$ whenever $(x_1, \dots, x_n) \in \mathcal{Z}$ has k elements in $\mathcal{X}$ and contains a suffix of length shorter than k belonging to $\mathcal{H}$.*

3. A conjugate prior for reversible random walks with amnesia

We will infer the transition probabilities of a reversible random walk with amnesia $(X_i)_{i\in\mathbb{N}}$ using the sequence of $\mathcal{X}$-valued states $(Z_1, \dots, Z_m)$. Our model, like the special case of a variable-order Markov chains, mitigates the need of estimating a large number of transition probabilities for dynamics with long memory in a subset of contexts. The major advantage of the approach we propose is that it does not require model selection or model averaging over the possible sets of histories $\mathcal{H}$. We propose to infer the law of a reversible sequence $(Z_1, Z_2 \dots)$ using a Bayesian model for random walks with amnesia. We specify a prior distribution that concentrates on amnesia processes with short memory. These are random walks with stationary measures w_n that tend to assign higher weights on the low memory states in $\mathcal{Z}$, such as $(s_{n-1}, \dots, s_{n-1}, x_n)$, compared to the high memory states, such as $(s_1, x_2 \dots, x_n)$, where $(x_2, \dots, x_n) \in \mathcal{X}^{n-1}$.

Let $w_{n+1}^0 : \mathcal{X}_s^{n+1} \to [0,\infty)$ be a balanced function, which determines the transition probabilities of a reversible recurrent random walk with amnesia. This function will specify a prior for w_{n+1}, the unknown parameter of the process $(X_i)_{i\in\mathbb{N}}$. We first define a reinforced process $(Y_i)_{i\in\mathbb{N}}$ parameterized by w_{n+1}^0. We then show it is

a mixture of reversible random walks with amnesia. The mixing distribution is a conjugate prior for estimating, given a path $X_1, \ldots, X_N$, with $N > n$, the transition probabilities of the nth-order random walk with amnesia. As in other reinforcement schemes, such as the Pólya urn and the edge-reinforced random walk (Diaconis and Rolles, 2006), the parameter of the prior, in our case w^0_{n+1}, and the unknown parameter w_{n+1} that we want to estimate, are functions defined on $\mathcal{X}_s^{n+1}$ which share two characteristics: they are both balanced and reversible.

Definition 2 *The process $(Y_i)_{i\in\mathbb{N}}$ takes values in $\mathcal{X}_s$. With probability 1, $(Y_1, \ldots, Y_n)$ is a palindrome, i.e. $(Y_1, \ldots, Y_n) = (Y_n, \ldots, Y_1)$, and $w^0_n(Y_1, \ldots, Y_n) > 0$. For $i > n$, the transition probabilities*

$$pr(Y_i|Y_1, \ldots, Y_{i-1}) = \frac{w^{i-n-1}_{n+1}(Y_{i-n}, \ldots, Y_i)}{\sum_{v\in\mathcal{X}_s} w^{i-n-1}_{n+1}(Y_{i-n}, \ldots, Y_{i-1}, v)}$$

are specified by the recursive reinforcement equations

$$\begin{aligned} w^{j+1}_{n+1}(u) = w^j_{n+1}(u) + & \\ \mathbf{1}(Y_{j+1} = u_1, \ldots, Y_{j+n+1} = u_{n+1}) + & \mathbf{1}(Y_{j+1} = u_{n+1}, \ldots, Y_{j+n+1} = u_1), \end{aligned} \tag{8}$$

where $j \geq 0$ and $u = (u_1, \ldots, u_{n+1}) \in \mathcal{X}_s^{n+1}$. The weights w^0_{n+1} parameterize the law of $(Y_i)_{i\in\mathbb{N}}$.

The reinforced processes $(Y_i)_{i\in\mathbb{N}}$ in Definition 2 constitutes a subclass of the *reinforced random walks with memory* introduced in Bacallado (2011) [Definition 3.1]. Each process within our subclass is identified by a reversible balanced function w^0_{n+1} which parametrizes an $\mathcal{X}_s$-valued reversible random walk with amnesia. Similarly, reinforced random walks with memory are defined using reversible balanced functions that parametrize reversible nth-order Markov chains. The reinforcement mechanisms in Definition 2 and in reinforced random walks with memory are identical. The next proposition shows that $(Y_i)_{i\in\mathbb{N}}$ can be used as Bayesian model to infer the transition probabilities of $(X_i)_{i\in\mathbb{N}}$. The proof of the proposition, together with a few additional remarks on the process $(Y_i)_{i\in\mathbb{N}}$, is included in in the Supplementary Materials.

Proposition 5 *The process $(Y_i)_{i\in\mathbb{N}}$ is a mixture of reversible random walks with amnesia. For every $N > n$*

$$pr(Y_1,\ldots,Y_N|Y_1,\ldots,Y_n) = \int \prod_{i=1}^{N-n} \frac{w_{n+1}(Y_i,\ldots,Y_{i+n})}{\sum_{v\in\mathcal{X}_s} w_{n+1}(Y_i,\ldots,Y_{i+n-1},v)} d\mu(w_{n+1}), \quad (9)$$

where μ is a distribution on the space of balanced functions that parametrize reversible random walks with amnesia.

This result relies on a de Finetti type representation theorem for Markov chains, which relies on the notion of Markov exchangeability (Diaconis and Freedman, 1980; Fortini and Petrone, 2014). We say a discrete process is Markov exchangeable if the probability of any finite path, can be expressed as a function of the initial state and the transition counts in the path between every pair of states in the state space. Important properties of $(Y_i)_{i\in\mathbb{N}}$, including both Markov exchangeability of the sequence of n-grams visited by the process and recurrence follow directly from the study of reinforced random walks with memory in Bacallado (2011). It is not hard to verify Markov exchangeability directly using a simple closed form expression for the conditional probabilities

$$\begin{aligned}\mathrm{pr}(Y_{n+1}=y_{n+1},\ldots,Y_N=y_N \mid Y_1=y_1,\ldots,Y_n=y_n) = \prod_i \mathbf{1}\Big[w^0_{n+1}(y_i,\ldots,y_{i+n})>0\Big]\times \\ \prod_{x:w^0_{n+1}(x)>0} g\Big[x,(y_1,\ldots,y_N)\Big] \times \prod_{x:w^0_n(x)>0} g'\Big[x,(y_1,\ldots,y_N)\Big],\end{aligned} \quad (10)$$

where, $(y_1,\ldots,y_n) = (y_n,\ldots,y_1)$, $w^0_n(y_1,\ldots,y_n) > 0$,

$$g\Big[x,(y_1,\ldots,y_N)\Big] = \begin{cases} \dfrac{\Gamma\Big(w^0_{n+1}(x)+\sum_i \mathbf{1}_x(y_i,\ldots,y_{i+n})+\mathbf{1}_x(y_{i+n},\ldots,y_i)\Big)^{1/2}}{\Gamma\Big(w^0_{n+1}(x)\Big)^{1/2}} & \text{if } (x_1,\ldots,x_{n+1}) \neq (x_{n+1},\ldots,x_1), \\ \dfrac{\Gamma\Big(w^0_{n+1}(x)/2+\sum_i \mathbf{1}_x(y_i,\ldots,y_{i+n})\Big)\times 2^{\sum_i \mathbf{1}_x(y_i,\ldots,y_{i+n})}}{\Gamma\Big(w^0_{n+1}(x)/2\Big)} & \text{if } (x_1,\ldots,x_{n+1}) = (x_{n+1},\ldots,x_1), \end{cases}$$

for every $x = (x_1, \ldots, x_{n+1}) \in \mathcal{X}_s^{n+1}$ that satisfies $w_{n+1}^0(x) > 0$, and

$$g'\Big[x, (y_1, \ldots, y_N)\Big] =$$
$$\begin{cases} \dfrac{\Gamma\Big(w_n^0(x)\Big)^{1/2}}{\Gamma\Big(w_n^0(x)+\sum_i \mathbf{1}_x(y_i,\ldots,y_{i+n-1})+\mathbf{1}_x(y_{i+n-1},\ldots,y_i)\Big)^{1/2}} & \text{if } (x_1,\ldots,x_n) \neq (x_n,\ldots,x_1), \\ \dfrac{\Gamma\Big([1-\mathbf{1}_x(y_1,\ldots,y_n)+w_n^0(x)]/2\Big)\times 2^{-\sum_i \mathbf{1}_x(y_i,\ldots,y_{i+n-1})}}{\Gamma\Big([1-\mathbf{1}_x(y_1,\ldots,y_n)+w_n^0(x)]/2+\sum_i \mathbf{1}_x(y_i,\ldots,y_{i+n-1})\Big)} & \text{if } (x_1,\ldots,x_n) = (x_n,\ldots,x_1), \end{cases}$$

for every $x = (x_1, \ldots, x_n) \in \mathcal{X}_s^n$ such that $w_n^0(x) = \sum_{x_{n+1}} w_{n+1}^0(x_1, \ldots, x_n, x_{n+1}) > 0$. The above expression specifies the law of $(Y_i)_{i\in\mathbb{N}}$ using the gamma function, $\Gamma : \mathbb{R}^+ \to \mathbb{R}^+$, and has similarities with other well known reinforcement schemes such as the edge-reinforced random walk (Diaconis and Rolles, 2006) and the Pólya urn model. The expression shows that, conditional on $(Y_1 = y_1, \ldots, Y_n = y_n)$, the probability of the event $(Y_1 = y_1, \ldots, Y_N = y_N)$, $N > n$, depends only on the counts

$$\sum_{i\leq N-n} \mathbf{1}_x(y_i, \ldots, y_{i+n});\ x \in \mathcal{X}_s^{n+1}.$$

Therefore, the sequence of n-grams $(Y_1, \ldots, Y_n), (Y_2, \ldots, Y_{n+1}), \ldots$, is Markov exchangeable.

In what follows we use the mixing distribution μ in proposition (5) as Bayesian prior for the unknown parameter w_{n+1} of our amnesia process $(X_i)_{i\in\mathbb{N}}$. Under this Bayesian model we can sample from the predictive distributions $\mathrm{pr}(Y_{N+1} \mid Y_1, \ldots Y_N)$. The computation of these predictive probabilties reduces to the evaluation of the reinforced weights w_{n+1}^N in Definition 2.

The fact that the initial n-gram in $(Y_i)_{i\in\mathbb{N}}$ is a palindrome is a necessary hypothesis for Proposition 5. However, we can define priors for a random walk with amnesia with initial n-gram $(y_1, \ldots, y_n) \neq (y_n, \ldots, y_1)$ by using the distribution of $(Y_i)_{i>m}$ conditional on $Y_1 = y_1, \ldots, Y_m = y_m$, with $(y_1, \ldots, y_m)$ specified upfront together with w_{n+1}^0.

Remark 1 *Consider a prior parameter w_{n+1}^0 with $w_{n+1}^0(x) > 0$ for every $x = (x_1, \ldots, x_{n+1})$ such that the transition from $(x_1, \ldots, x_n) \in \mathcal{X}_s^n$ to x_{n+1} is allowed by the definition (1) of random walk with amnesia. Assume that the investigator*

observes the trajectory $(X_1, \ldots, X_s)$ *of a recurrent random walk with amnesia with unknown transition probabilities, and fix a recurrent n-gram* $(x_1, \ldots, x_n) \in \mathcal{X}_s^n$. *The linear reinforcement (8) indicates that, the Bayesian estimate, of the transition probability* $\frac{w_{n+1}(x_1,\ldots,x_{n+1})}{\sum_v w_{n+1}(x_1,\ldots,x_n,v)}$, *defined as the posterior mean, converges in probability to the empirical estimate*

$$\frac{\sum_{j=1}^{s} \mathbf{1}(X_j = x_1, \ldots, X_{j+n} = x_{n+1}) + \mathbf{1}(X_j = x_{n+1}, \ldots, X_{j+n} = x_1)}{\sum_{j=1}^{s} \mathbf{1}(X_j = x_1, \ldots, X_{j+n-1} = x_n) + \mathbf{1}(X_j = x_n, \ldots, X_{j+n-1} = x_1)}$$

when the the length of the trajectory s diverges. This fact follows from the convergence in probability, for any $(x_1, \ldots, x_n) \in \mathcal{X}_s^n$ *and* $(x_1, \ldots, x_{n+1}) \in \mathcal{X}_s^{n+1}$, *of the ratios*

$$\frac{\sum_{j=1}^{s-n} \mathbf{1}(X_j = x_1, \ldots, X_{j+n} = x_{n+1})}{s}, \quad \textit{and} \quad \frac{\sum_{j=1}^{s-n+1} \mathbf{1}(X_j = x_1, \ldots, X_{j+n} = x_n)}{s}$$

to the parameters $w_{n+1}(x_1, \ldots, x_{n+1})$ *and* $w_{n+1}(x_1, \ldots, x_n)$ *respectively.*

4. Posterior simulations

The process $(Y_i)_{i\in\mathbb{N}}$ in Definition 2 is used as a prior distribution for the random walk with amnesia. This process is driven by a random nth-order transition matrix (Proposition 5). Recall that the observable sequence $(Z_j)_{j\in\mathbb{N}}$ takes values in $\mathcal{X}$. Let $Z_j = Y_{\tau_j}$, where $\tau_j = \min\{\tau; \tau > \tau_{j-1}, (Y_{\tau-n}, ..., Y_{\tau-1}) \in \mathcal{Z}, Y_\tau \in \mathcal{X}\}$ and $\tau_0 = n$. The length of memory of the process $(Z_j)_{j\in\mathbb{N}}$ at each transition is captured by a sequence of latent variables $T_j := \max\{t : \tau_j - t > 0, (Y_{\tau_j-1}, \ldots, Y_{\tau_j-t}) \in \mathcal{X}^t\}$, $j \geq 1$. Without loss of generality, assume $Y_1 = \ldots = Y_n = s_n$.

The observed states $Z_1, \ldots, Z_{r-1}$ and the lengths of memory $T_1, \ldots, T_{r-1}$ identify the path $Y_1, \ldots, Y_{\tau_{r-1}}$, which in turn can be used to obtain the reinforced weights $\{w_{n+1}^{\tau_{r-1}-n}(u); u \in \mathcal{X}_s^{n+1}\}$ and to compute the predictive distribution

$$\text{pr}(Y_{1+\tau_{r-1}}, \ldots, Y_{\tau_r} \mid T_1, \ldots, T_{r-1}, Z_1, \ldots, Z_{r-1}) = \text{pr}(Y_{1+\tau_{r-1}}, \ldots, Y_{\tau_r} \mid Y_1, \ldots, Y_{\tau_{r-1}}).$$

This makes it straightforward to sample T_r conditional on $(T_1, \ldots, T_{r-1}, Z_1, \ldots, Z_r)$, which will be useful in the sequential importance sampling algorithm (Gordon et al., 1993) proposed below.

The goal of the algorithm is to infer the lengths of memory $T_1, \ldots, T_m$ given a sequence of observations $Z_1, \ldots, Z_m$. The algorithm is initialized with particles

$t_1^{(i)} = 0$ and importance weights $\mathbf{v}_1^{(i)} = 1$ for all $i = 1, \ldots, N$. For each $r = 2, \ldots, m$, the particle $t_r^{(i)}$ is sampled from the conditional distribution of T_r given the observed random variables $(Z_1 = z_1, \ldots, Z_r = z_r)$ and $(T_1 = t_1^{(i)}, \ldots, T_{r-1} = t_{r-1}^{(i)})$ for all $i = 1, \ldots, N$. At each step, the importance weights are updated by

$$\mathbf{v}_r^{(i)} = \mathbf{v}_{r-1}^{(i)} \mathrm{pr}(Z_r = z_r \mid Z_1 = z_1, \ldots, Z_{r-1} = z_{r-1}, T_1 = t_1^{(i)}, \ldots, T_{r-1} = t_{r-1}^{(i)}).$$

Finally, we approximate the posterior distribution of the lengths of memory using the weighted particles,

$$\mathrm{pr}(T_1, \ldots, T_m \mid Z_1, \ldots, Z_m) \approx \sum_{i=1}^{N} \tilde{\mathbf{v}}_m^{(i)} \mathbf{1}(T_1 = t_1^{(i)}, \ldots, T_m = t_m^{(i)}), \tag{11}$$

where

$$\tilde{\mathbf{v}}_r^{(i)} = \frac{\mathbf{v}_r^{(i)}}{\sum_{j=1}^{N} \mathbf{v}_r^{(j)}}.$$

It is well understood that in many cases, a better approximation is obtained by combining the structure of sequential importance sampling with resampling operations (Gordon et al., 1993). We will apply a simple strategy known as sequential importance resampling, or the bootstrap particle filter. A resampling operation at time r consists of the following two steps. First, after the weights $\tilde{\mathbf{v}}_r^{(i)}$, $i = 1, \ldots, N$, have been computed, N random variables a_i, $i = 1, \ldots, N$, taking values in $\{1, \ldots, N\}$ are independentely generated with $\mathrm{pr}(a_i = j) = \tilde{\mathbf{v}}_r^{(j)}$, $j = 1, \ldots, N$. Second, each particle $(t_1^{(i)}, \ldots, t_r^{(i)})$ is replaced by $(t_1^{(a_i)}, \ldots, t_r^{(a_i)})$, and the weights $\mathbf{v}_r^{(i)}$ are set equal to 1, for $i = 1, \ldots, N$. The algorithm is then ready to proceed as described in the previous paragraph.

Let $r_1, \ldots, r_k$ be the times at which resampling steps are made, and let $\hat{\mathbf{v}}_r$ be the average of $\mathbf{v}_r^{(1)}, \ldots, \mathbf{v}_r^{(N)}$ before the resampling operation at time r. Then, it can be shown that $\prod_{j=1}^{k} \hat{\mathbf{v}}_{r_j}$ is an unbiased estimate for the marginal probability of the observation $\mathrm{pr}(Z_1, \ldots, Z_{r_k})$; see Proposition 7.4.1 in del Moral (2004). This will be convenient in the following sections for estimating predictive probabilities of the form $\mathrm{pr}(Z_{r+1}, \ldots, Z_m \mid Z_1, \ldots, Z_r)$.

The particles $(t_1^{(i)}, \ldots, t_m^{(i)}; i = 1, \ldots, N)$ approximate the posterior distribution of the length of memories $T_1, \ldots, T_m$. They can be used to approximately sample

from the predictive distribution of the process $(Z_i)_{i\in\mathbb{N}}$. A long simulation from the predictive distribution yields, by de Finetti representation (Proposition 5) arguments, an approximate posterior sample of the transition probabilities that drive the random walk with amnesia. Posterior samples of w_{n+1} can be used, among other things, to calculate predictive probabilities of the form $\mathrm{pr}(Z_{r+1}, \ldots, Z_m \mid Z_1, \ldots, Z_r)$ via a forward-backward algorithm; this is an alternative to the estimates mentioned above.

5. Simulation Study

We specify the prior model choosing, for every $(x_1, \ldots, x_n) \in \mathcal{Z}$,

$$w^0_{n+1}(x_1, \ldots, x_{n+1}) = c \times (|\mathcal{X}|b)^{-\sum_i \mathbf{1}(x_i \in \mathcal{X})}, \tag{12}$$

for some $b, c > 0$, if the random walk with amnesia allows the transition $(x_1, \ldots, x_n) \to x_{n+1}$ and zero otherwise. This reduces the set of hyperparameters that we need to tune to the pair (b, c). The initial function w^0_{n+1} corresponds to a random walker that, from any n-gram in $\mathcal{Z}$, with some fixed probabilty forget the first $\mathcal{X}$-valued element in his memory. Also, under w^0_{n+1}, transitions to the $\mathcal{X}$-valued states from any n-gram in $\mathcal{Z}$ are all equally likely.

The parameter b tunes the length of memory. The greater b, the shorter the length of memory tends to be during the reinforced walk $(Y_i)_{i\geq 1}$. The parameter c has an interpretation similar to the total initial mass of the the edge-weighted graph in Diaconis and Rolles (2006) and tunes the concentration of w_{n+1} around the prior mean. Figure 1b illustrates how the choice of the parameters (b, c) tunes the complexity of a typical random walk with amnesia with 10 states sampled from the prior. As a measure of complexity, we use the entropy rate $\lim_{n\to\infty} n^{-1} H(X_1, \ldots, X_n)$ of the random walk with amnesia, where H is the Shannon entropy.

We tested the proposed model using data simulated from a reversible fourth order Markov chain in the space of nucleic bases $\{a, c, g, t\}$. The transition probabilities of the Markov chain were the frequencies of every fourth order transition in a cyclic DNA sequence of 37,243 bases read in both directions (RefSeq NC 021042). The random walk with amnesia was fit to a simulated sequence by sequential importance resampling, and the predictive probability of a different simulated sequence was approximated using the importance weights as discussed at the end of Section 4.

We used $N = 1{,}000$ particles, and performed resampling operations every 3 steps. Figure 2 shows the log-predictive probabilities per step given training sequences of lengths ranging between 100 and 1,900. The hyperparameters b and c were selected to optimize the probability of the training sequence provided by the sequential importance sampling algorithm over a grid with $b \in \{0.5, 1, 1.5, 2\}$ and $c \in \{0.5, 1, 1.5, 2\}$.

For the purpose of comparison, Figure 2 shows the same statistics for a Bayesian analysis of first order and fourth order reversible Markov models (Bacallado, 2011). For each model, the concentration parameter, i.e. the initial weight of the palindrome in the model of Bacallado (2011), is optimized to maximize the probability of the training sequence. Even though the sequence was generated from a fourth order Markov chain, the fourth order model performs poorly compared to the random walk with amnesia. We observe, as expected, that the inferred random walk with amnesia adapts to the complexity of the training sequence, with longer lengths of memory associated with the more frequent contexts.

Bacallado (2011) also defines a prior for reversible variable-order Markov models. We compared the random walk with amnesia to a Bayesian analysis based on variable-order Markov models, with and without reversibility. In the non-reversible case, we apply an independent Dirichlet prior to the transition probabilities out of every context. Before applying these models, it is necessary to estimate the appropriate lengths of memory. For this purpose, we applied a context tree pruning algorithm similar in structure to the Context algorithm of Rissanen (1983).

We describe the main characteristics of the pruning algorithm. First, the algorithm grows the largest possible context tree in which every context appears at least 5 times in the training sequence and is no longer than 6. Then, we employ a backward tree-pruning procedure with a local criterion. The criterion is a Bayes factor comparing the variable-order model before and after pruning the tree. We loop through the leaves in alphabetical order; in the Context algorithm, which uses a slightly different local criterion, the order is irrelevant. The threshold for the Bayes factor, and the concentration parameter of the prior for the variable-order Markov models are used as tuning parameters and selected to optimize the probabil-

ity of the training data under the model. The pruning strategy is slightly modified in the reversible case, as the set of contexts must satisfy the closure properties in Proposition 4.2 of Bacallado (2011).

Figure 2 shows that, as expected, the reversible models outperform non-reversible models. The random walk with amnesia outperforms both of the model selection schemes in this example.

To evaluate the performance of the methods above on data which is not reversible, we repeated the analysis using the original cyclic DNA sequence instead of the reversible simulated data. Figure 3 shows that, as one would expect, models that enforce reversibility pay a price in bias and perform worse than an appropriately pruned variable-order Markov model for longer training sets. This is not the case when reversibility is a given, as in the application of the following section.

6. Molecular dynamics

We analyze a molecular dynamics simulation of a protein known as the WW domain (Shaw et al., 2010). Markov models have become, in recent years, essential tools for the analysis of this type of dynamical data (Prinz et al., 2011). The data consist of a sequence of states representative of distinct conformations of the protein observed at regular time intervals. In this application, the stochastic law of the process is reversible due to the nature of Hamiltonian mechanics. We have 10 states in $\mathcal{X}$ and a trajectory of length 10,000, with one observation every 20 nanoseconds. We use the first half of the trajectory as a training dataset and the second as a test dataset.

Figure 1a plots the trajectory. The plot shows that the process alternates phases during which it remains stable at state 0 with phases during which it rapidly moves across states. These phases correspond to periods during which the protein is folded and unfolded, respectively.

We trained a reversible random walk with amnesia with $n = 6$ by sequential importance resampling. We used $N = 1{,}000$ particles and performed resampling operations every 3 steps. The hyperparameters b and c were chosen to maximize the probability of the training sequence under the model. Table 1 shows the selected parameters and the loglikelihood of the test sequence. To evaluate the efficacy of the sequential importance resampling algorithm, we replicated the computations

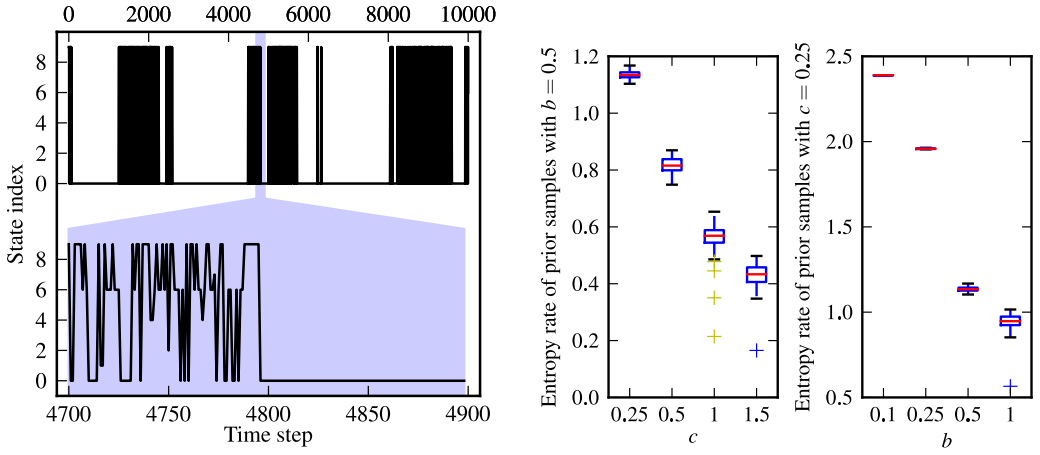

Fig. 1. *Left:* Conformational states of the WW domain, indexed by $0, \ldots, 9$, observed in a molecular dynamics simulation. The top panel shows the full sequence of 10,000 steps, while the bottom panel shows a magnified subsequence. *Right:* The entropy rates $\lim_{n\to\infty} n^{-1} H(X_1, \ldots, X_n)$ of 100 random walks with amnesia sampled from the prior distribution under eight choices of (b, c).

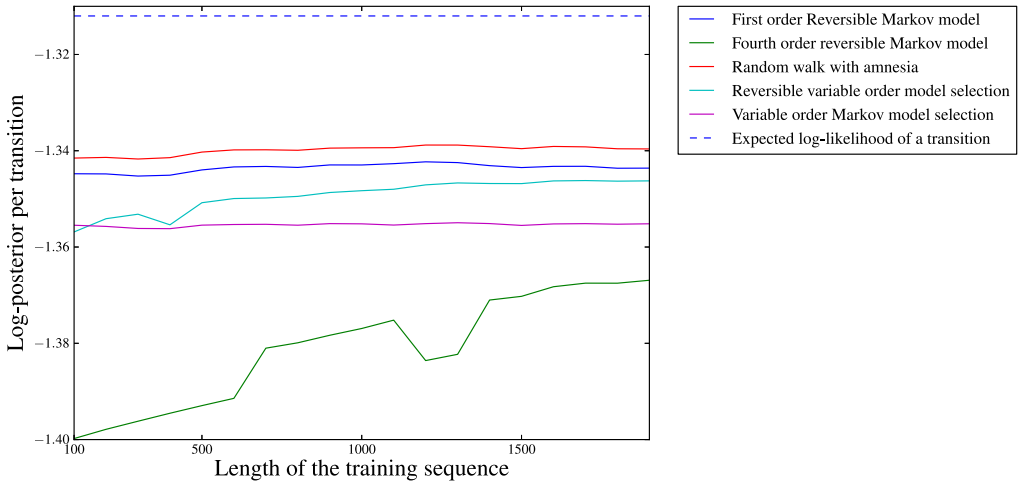

Fig. 2. Reversible simulation example. Log-posterior predictive probabilities per transition for a left-out test set, as a function of the length of the training sequence. The dashed line shows the expected loglikelihood per step in the model used to simulate the sequence.

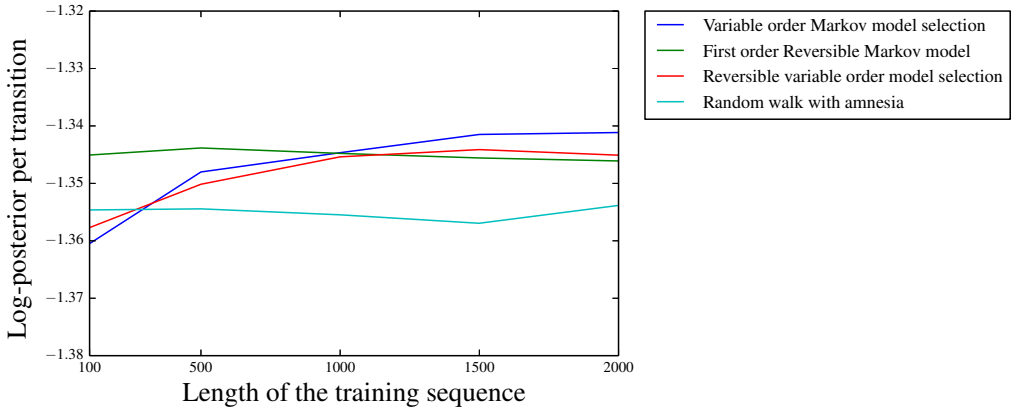

Fig. 3. Non-reversible example. Log-posterior predictive probabilities per transition for a left-out test set, as a function of the length of the training sequence.

10 times at every hyperparameter setting. The unbiased estimates of marginal likelihood varied little in these iterations; for example, the estimate shown in the last row of table 1 had a standard deviation of 67.

To compare our procedure with some common alternatives, we considered five methods reviewed by Begleiter et al. (2004), none of which takes into account reversibility. Following the recommendations in Begleiter et al. (2004), we tuned the parameters of each method via 3-fold cross-validation using only the training sequence. We then computed the loglikelihood of the test dataset at the optimal parameters. Table 1 lists the methodologies, the tuning parameters selected, in the notation of Begleiter et al. (2004), and the loglikelihood of the test trajectory. The best of the methods in this example is Decomposed Context Tree Weighting. The random walk with amnesia has a higher loglikelihood than all five methods.

Finally, we applied the Bayesian analyses of variable-order Markov models described in Section 6, selecting the model selection threshold t and the concentration parameter c which optimize the likelihood of the training data.

The results are summarized in Table 1. The reversible model outperforms the non-reversible model, but both of the methods are outperformed by Decomposed Context Tree Weighting and the random walk with amnesia. The tuning of the

Table 1. Predictive performance of the random walk with amnesia compared to the methods reviewed in Begleiter et al. (2004) and two model selection approaches for variable-order Markov models. The tuning parameters of each method from Begleiter et al. (2004) were chosen via cross-validation using 50% of the data. We report the range of values explored for each tuning parameter and the optimal value in bold. We also report the corresponding loglikelihood of the test dataset.

Method	Test loglikelihood	Tuning parameters
Decomposed Context Tree Weighting	-3210.4	$D \in \{1, 3, 5, 7, 9, \mathbf{15}, 20\}$
Prediction by Partial Matching	-3359.8	$D \in \{\mathbf{1}, 3, 5, 7, 9, 15, 20\}$
Lempel Ziv 78	-4041.4	None
Lempel Ziv MS	-3862.1	$M \in \{\mathbf{0}, 2, 4, 6, 8\}$ $S \in \{0, 2, 4, \mathbf{6}, 8\}$
Probabilistic Suffix Trees	-3280.9	$P_{\min} \in \{0.0001, \mathbf{0.001}, 0.01, 0.1\}$ $\gamma \in \{0.0001, \mathbf{0.001}, 0.01, 0.1\}$ $D \in \{\mathbf{1}, 3, 5, 7, 9, 15, 20\}$, $r = 1.05$, $\alpha = 0$
Variable-order Markov model	-3308.4	$c \in \{0.0001, \mathbf{0.001}, 0.01, 0.1, 1\}$, $t \in \{-10, \mathbf{0}, 10\}$
Reversible variable-order Markov model	-3268.8	$c \in \{0.0001, \mathbf{0.001}, 0.01, 0.1, 1\}$, $t \in \{-10, \mathbf{0}, 10\}$
Random walk with amnesia	-3015.4	$b \in \{0.1, 0.4, \mathbf{0.6}, 1, 1.5\}$, $c \in \{\mathbf{0.5}, 1, 1.5, 2\}$

random walk with amnesia was repeated using a 3-fold cross-validation strategy, and this had a negligible effect on the results in Table 1.

References

Annis, D., Kiessler, P., Lund, R., and Steuber, T. (2010). Estimation in reversible markov chains. *Am. Stat.,* **64**, 116–120.

Bacallado, S. (2011). Bayesian analysis of variable-order, reversible markov chains. *Ann. Stat.,* **39**, 838–864.

Begleiter, R., El-Yaniv, R., and Yona, G. (2004). On prediction using variable order Markov models. *J. Artif. Intell. Res.,* **22**, 385–421.

Bühlmann, P. (2000). Model selection for variable length Markov chains and tuning the context algorithm. *Ann. I. Stat. Math.,* **52**, 287–315.

Bühlmann, P. and Wyner, A. J. (1999). Variable length Markov chains. *Ann. Stat.,* **27**, 480–513.

Cappé, O., Moulines, E., and Ryden, T. (2005). *Inference in Hidden Markov Models.* Springer.

del Moral, P. (2004). *Feynman-Kac Formulae.* Springer.

Diaconis, P. and Freedman, D. (1980). de Finetti's theorem for Markov chains. *Ann. Prob.,* **8**, 115–130.

Diaconis, P. and Rolles, S. (2006). Bayesian analysis for reversible Markov chains. *Ann. Stat.,* **34**, 1270–1292.

Fortini, S. and Petrone, S. (2014). Predictive characterizations of mixtures of Markov chains. *arXiv:1406.5421.*

Gordon, N. J., Salmond, D. J., Smith, A. F. M. & Steuber, T. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proc. Rad. Sig. Pr.* **140**, 107–113.

Kelly, F. P. (1979). *Reversibility and stochastic networks.* Chichester: Wiley.

MacKay, D. and Peto, L. (1995). A hierarchical Dirichlet language model. *Nat. Lang. Eng.,* **1**, 289–308.

Mochihashi, D. and Sumita, E. (2008). The infinite Markov model. *Adv. Neural Inf. Process. Syst.,* **20**, 1017–1024.

Palla, K., Knowles, D., and Ghahramani, Z. (2008). A reversible infinite HMM using normalised random measures. In *Proceedings of the 31st International Conference on Machine Learning.* International Machine Learning Society.

Prinz, J., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J., Schütte, C., and Noé, F. (2011). Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.,* **134**, 174105.

Rissanen, J. (1983). A universal data compression system. *IEEE T. Inform. Theory.,* **29**, 656–664.

Ron, D., Singer, Y., and Tishby, N. (1996). The power of amnesia: Learning probabilistic automata with variable memory length. *Mach. Learn.,* **25**, 117–149.

Shaw, D., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R., Eastwood, M., Bank, J., Jumper, J., Salmon, J., Shan, Y., et al. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science,* **330,** 341–346.

Teh, Y. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics*, pp. 985–992. Association for Computational Linguistics.

Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry.* North-Holland personal library. Amsterdam: North-Holland.

Weinberger, M., Rissanen, J., and Feder, M. (1995). A universal finite memory source. *IEEE T. Inform. Theory.,* **41**, 643–652.

Wood, F., Archambeau, C., Gasthaus, J., James, L. F., and Teh, Y. W. (2009). A stochastic memorizer for sequence data. In *Proceedings of the 26th International Conference on Machine Learning,* pp. 1129-1136. International Machine Learning Society.