

Sparse Markov Chains for Sequence Data[‡]

VÄINÖ JÄÄSKINEN and JIE XIONG

Department of Mathematics and Statistics, University of Helsinki

JUKKA CORANDER

Department of Mathematics and Statistics, University of Helsinki

Department of Mathematics, Åbo Akademi University

TIMO KOSKI

Department of Mathematics, KTH Royal Institute of Technology

ABSTRACT. Finite memory sources and variable-length Markov chains have recently gained popularity in data compression and mining, in particular, for applications in bioinformatics and language modelling. Here, we consider denser data compression and prediction with a family of sparse Bayesian predictive models for Markov chains in finite state spaces. Our approach lumps transition probabilities into classes composed of invariant probabilities, such that the resulting models need not have a hierarchical structure as in context tree-based approaches. This can lead to a substantially higher rate of data compression, and such non-hierarchical sparse models can be motivated for instance by data dependence structures existing in the bioinformatics context. We describe a Bayesian inference algorithm for learning sparse Markov models through clustering of transition probabilities. Experiments with DNA sequence and protein data show that our approach is competitive in both prediction and classification when compared with several alternative methods on the basis of variable memory length.

Key words: Bayesian learning, data compression, predictive inference, Markov chains, variable order Markov models

1. Introduction

Variable order Markov (VOM) chain models pioneered by Rissanen (1983) have been a subject of intensive further research in the past two decades (Weinberger *et al.*, 1995; Bühlmann and Wyner, 1999; Bacallado, 2011). Compression of data sequences using such models can be understood as a characterization of the observed information in terms of the learned generating model and the rate of compression as proportional to inverse of the parametric dimension of the learned model. Because the variable order models allow for a substantially higher rate of data compression than ordinary Markov chain (MC) models, they have recently become popular for various applications in modelling DNA data (Ben-Gal *et al.*, 2005; Zhang *et al.*, 2005; Browning, 2006; Corander *et al.*, 2009). Another important application of variable order memory is modelling of natural languages (Wood *et al.*, 2009; Gasthaus *et al.*, 2010). For an early Bayesian approach to predictive language modelling based on Markov-type dependence and a hierarchical Dirichlet prior, see also MacKay & Peto (1995). For a review and comparison of several algorithms for data compression with VOM models, see Begleiter *et al.* (2004). Roos & Yu (2009) introduced a method for estimating context tree models in binary state spaces

[‡] This article was published online on [31 October 2013]. The images of figures 1 and 3 were subsequently changed. This notice is included in the online and print versions to indicate that both have been corrected [27 December 2013].

based on a transformation leading to logistic regression models for which Lasso penalization is applied.

Here, we consider compression and prediction of data sequences using Bayesian inference on sparse MC (SMC) models, which need not correspond to a hierarchical representation of contexts used in variable order and variable length MCs (VLMCs). We show that the SMC models can lead to significantly improved predictions and higher rate of data compression than VOM and VLMC models. Related classes of Markov models were introduced in García & González-López (2010), who derived an asymptotic criterion for learning of partition models, and in Farcomeni (2011), who considered hidden Markov partition models and proposed learning by an expectation–maximization algorithm (EM), where the number of hidden states is fixed.

Bayesian inference for VOM models seems to have gained attention only very recently. Dimitrakakis (2010a) considered Bayesian estimation of VOM models using an approach similar to the context tree weighting (Willems *et al.*, 1995), whereas Bacallado (2011) introduced a conjugate prior for reversible VLMC models. García & González-López (2010) derived a consistent information theoretic criterion for model comparison based on an asymptotic expansion generalizing the results presented in Csiszár & Shields (2000) and Csiszár & Talata (2006). Using the results derived in Corander *et al.* (2009) for VLMC models, we introduce both prior predictive and posterior predictive distributions for sparse Markov models to enable Bayesian inference within this class of models. It is shown that the inference problem can be formulated as a clustering of populations (contexts/words) for which the data correspond to observed counts of transitions to the underlying alphabet.

The outline of the paper is as follows. In Section 2, we introduce the SMC models and investigate their relation to VLMC models. A Bayesian inference method for SMC models is described in Section 3, together with several predictive comparisons with other variable order methods. Some remarks about possible generalization of sparse models, in particular, in the context of bioinformatics applications are given in the final section.

2. Sparse Markov models and variable length memory

Let $\mathcal{X} = \{1, \dots, J\}$ be a finite alphabet and X_0, X_1, \dots, X_n a sequence of random variables that take values in \mathcal{X} . Assume that the sequence can be represented by a time homogeneous MC $\{X_n\}_{n=0}^\infty$ of a finite order m . To enable comparison of alternative model definitions, we let \mathcal{X}^m be the enlarged alphabet with $|\mathcal{X}^m| = J^m$ values. An arbitrary time homogeneous MC $\{X_n\}_{n=0}^\infty$ of a finite order m can always be represented by transforming it to a first-order MC $\{Z_n\}_{n=0}^\infty$ with transition probabilities

$$P = (p_{i|j})_{i,j=1}^{J^*, J^*}, \quad (1)$$

where i, j are arbitrary values in \mathcal{X}^m and each row has exactly J non-zero transition probabilities from the state i to state j in the vector $\mathbf{p}_{i|} = (p_{i|j})_{j=1}^{J^*}$. Note that for typographical clarity, in all the examples in the succeeding text, we use upper case letters instead of $p_{i|j}$ to denote the transition probabilities when the states involved are specified explicitly.

The following definition specifies a class of Markov models on the basis of partitions of \mathcal{X}^m .

Definition 1. Sparse Markov chain (SMC). Let $\{X_n\}_{n=0}^\infty$ be a time homogeneous MC of a finite order m transformed to a first-order MC $\{Z_n\}_{n=0}^\infty$. Let $S = (s_1, \dots, s_k)$ be a partition of \mathcal{X}^m such that the transition probability vectors satisfy the equality $p_{i|} = p_{j|}$ for all pairs of states $\{i, j\} \in s_c, c = 1, \dots, k$, and \mathcal{P} the corresponding set of k transition probability distributions in \mathcal{X}^m . If $k < |\mathcal{X}^m|$, the pair (S, \mathcal{P}) is called an SMC (of order m).

An SMC model is a special case of an MC, where two or more transition probability vectors have identical values, such that the effective dimension of the parameter space is reduced, which is illustrated by the following examples.

Example 1. Consider a second-order MC with the state space $\mathcal{X} = \{A, C, G, T\}$ and define a partition S according to

$$S = \{\{AA, AC, AG, AT\}, \{CA, CC, CG, CT\}, \{GA, GC, GG, GT\}, \{TA, TC, TG, TT\}\}.$$

Using the transformation of the state space to obtain a representation as a first-order MC, this partition corresponds to the four transition probability distributions

$$\theta^{(1)} = P_{AA|\cdot} = P_{AC|\cdot} = P_{AG|\cdot} = P_{AT|\cdot}$$

$$\theta^{(2)} = P_{CA|\cdot} = P_{CC|\cdot} = P_{CG|\cdot} = P_{CT|\cdot}$$

$$\theta^{(3)} = P_{GA|\cdot} = P_{GC|\cdot} = P_{GG|\cdot} = P_{GT|\cdot}$$

$$\theta^{(4)} = P_{TA|\cdot} = P_{TC|\cdot} = P_{TG|\cdot} = P_{TT|\cdot}$$

which imply that for each X_n , the preceding state X_{n-1} is irrelevant for predicting the state of X_n , whereas X_{n-2} is always relevant, whatever the state of X_{n-2} is. The likelihood of the data sequence $(x_0x_1 \cdots x_8) = (AAGTCCAAA)$ then equals

$$P_{AA} \cdot \theta_{AG}^{(1)} \cdot \theta_{GT}^{(1)} \cdot \theta_{TC}^{(3)} \cdot \theta_{CC}^{(4)} \cdot \theta_{CA}^{(2)} \cdot \theta_{AA}^{(2)} \cdot \theta_{AA}^{(1)}, \quad (2)$$

where the probability of the initial state P_{AA} is considered fixed and the elements of the transition probability vectors are indexed by the target states, so that, for instance, $\theta_{CA}^{(2)} = P_{CC|CA}$ and $\theta_{AA}^{(2)} = P_{CA|AA}$. Ignoring the probability distribution of the initial state as customary in likelihood inference for MCs, the SMC model based on the aforementioned partition reduces the number of free parameters from 48 to 12. The transition probabilities for the ordinary MC in the transformed state space are shown in the succeeding text, using lower case letters to obtain a more compact notation in which the conditional distribution of the current state given an arbitrary previous state is denoted by $p_{\cdot|\cdot}$.

$X_{n-1} \backslash X_n$	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0	0	0	0	0	0	0	0	0
AC	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0	0	0	0	0
AG	0	0	0	0	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0
AT	0	0	0	0	0	0	0	0	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$
CA	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0	0	0	0	0	0	0	0	0
CC	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0	0	0	0	0
CG	0	0	0	0	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0
CT	0	0	0	0	0	0	0	0	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$
GA	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0	0	0	0	0	0	0	0	0
GC	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0	0	0	0	0
GG	0	0	0	0	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0
GT	0	0	0	0	0	0	0	0	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$
TA	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0	0	0	0	0	0	0	0	0
TC	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0	0	0	0	0
TG	0	0	0	0	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	0	0	0	0
TT	0	0	0	0	0	0	0	0	0	0	0	0	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$	$p_{\cdot \cdot}$

Example 2. Consider again a second-order MC with the state space $\mathcal{X} = \{A, C, G, T\}$ and define the following partition of \mathcal{X}^m with 12 classes:

$$S = \{\{AA, CA, GA, TA\}, \{GT, TT\}, \{AC\}, \{AG\}, \{AT\}, \{CC\}, \\ \{CG\}, \{CT\}, \{GC\}, \{GG\}, \{TC\}, \{TG\}\}.$$

The resulting parametric sparsity is more modest (from 48 to 36 free parameters), such that the transition probability vectors corresponding to the two foremost classes in S are defined as

$$\theta^{(1)} = P_{AA|\cdot} = P_{CA|\cdot} = P_{GA|\cdot} = P_{TA|\cdot} \\ \theta^{(2)} = P_{GT|\cdot} = P_{TT|\cdot}.$$

Each of the remaining classes contains only a single state, which does not reflect any additional sparsity, and the corresponding probabilities are explicitly denoted by the pair of states according to the particular transition. Here, the likelihood of the previously introduced data sequence $(x_0 x_1 \dots x_8) = (AAGTCCAAA)$ can be written as

$$P_{AA} \cdot \theta_{AG}^{(1)} \cdot P_{AG|GT} \cdot \theta_{TC}^{(2)} \cdot P_{TC|CC} \cdot P_{CC|CA} \cdot \theta_{AA}^{(1)} \cdot \theta_{AA}^{(1)}, \quad (3)$$

where again the probability of the initial state P_{AA} is considered fixed.

The partitions in the two examples of SMC models enjoy different properties in terms of possibility to derive an alternative representation of the likelihood using so-called contexts (Rissanen, 1983; Bühlmann and Wyner, 1999), where for each X_n , the finite history X_{n-m}, \dots, X_{n-1} is mapped to the shortest possible context dependent subset X_{n-r}, \dots, X_{n-1} , $r \leq m$, such that

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_{n-m} = x_{n-m}) \\ = P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_{n-r} = x_{n-r}); \quad (4)$$

that is, the history is generally truncated to an order, which varies, depending on the actual outcome $X_{n-1} = x_{n-1}, \dots, X_{n-m} = x_{n-m}$. We make the difference between models in Examples 1 and 2 more explicit after introducing some additional notation.

Definition 2. Let $\{X_n\}_{n=0}^\infty$ be a time homogeneous MC of a finite order m . Let $\mathcal{B} = \{B_c \subset \mathcal{X}^{r_c} : c = 1, \dots, k\}$ be the set of k contexts for a VLHC defined by its context function f , where $r_c \leq m$ is the length of a particular context. When the cardinality $|B_c| = 1$, then B_c consists of the single $b^{(r)}$, $1 \leq r \leq m$, for which all paths x_{n-1}, \dots, x_{n-m} with $b^{(r)}$ as suffix satisfy the constraint

$$P(x_n = j | b^{(r)}, x_{n-r-1}, \dots, x_{n-m}) = P(x_n = j | b^{(r)}).$$

When the cardinality $1 < |B_c| < |\mathcal{X}|$, then B_c is a set of strings with a common suffix $b^{(r)}$, $1 \leq r \leq m$, such that for all paths x_{n-1}, \dots, x_{n-m} with $b^{(r)}$ as suffix and with $x_{n-r-1} \in A_{b^{(r)}} \subset \mathcal{X}$ satisfy the constraint

$$P(x_n = j | b^{(r)}, x_{n-r-1} \in A_{b^{(r)}}, \dots, x_{n-m}) = P(x_n = j | b^{(r)} [A_{b^{(r)}}]),$$

where $b^{(r)} [A_{b^{(r)}}]$ indicates the suffix $b^{(r)}$ concatenated by any letter in $A_{b^{(r)}}$. Let $\mathcal{P} = \{p_{B_c} : c = 1, \dots, k\}$ be the set of k conditional distributions associated with the elements of \mathcal{B} , where each $p_{B_c} = (p_{B_c|j})_{j=1}^J$, with $p_{B_c|j}$ either equal to $P(x_n = j | b^{(r)})$ or to $P(x_n = j | b^{(r)} [A_{b^{(r)}}])$, depending on the cardinality of B_c .

Note that the restriction $A_{b^{(r)}} \subset \mathcal{X}$ simply states that $A_{b^{(r)}}$ must be a proper subset of the alphabet \mathcal{X} , otherwise B_c could be reduced to a set of cardinality one, because then all preceding states x_{n-r-1} lead to the same conditional probability distribution that depends only on $b^{(r)}$. Hence, when B_c is a set with cardinality larger than one, it represents the set of contexts sharing a common suffix such that the corresponding transition probabilities are equal. The notation $b^{(r)} [A_{b^{(r)}}]$ is inspired by Mächler & Bühlmann (2004), who explicitly considered this more flexible and parsimonious class of VLMC models where a context can be a set instead of just a single string.

Example 3. Re-consider the SMC model from Example 2. Define the VLMC model with the contexts $\mathcal{B} = \{A, [GT]T, AC, CC, GC, TC, AG, CG, GG, TG, AT, CT\}$, written in the reverse order from left to right. Here, there is a single context set with cardinality larger than one, where $A_{b^{(r)}} = \{G, T\}$ and $b^{(r)} = T$, which implies that the equality (4) holds for the two pairs of states $(x_{n-1} = T, x_{n-2} = G), (x_{n-1} = T, x_{n-2} = T)$, such that

$$\begin{aligned} P(x_n = j | x_{n-1} = T, x_{n-2} = G) &= P(x_n = j | x_{n-1} = T, x_{n-2} = T) \\ &= P(x_n = j | b^{(r)} [A_{b^{(r)}}]). \end{aligned} \quad (5)$$

As can be seen from the set of contexts \mathcal{B} , for this model, it is necessary to specify 12 transition probability vectors, instead of the 16 required by a full second-order MC. For the data sequence $(x_0 x_1 \cdots x_8) = (AAGTCCAAA)$, the likelihood defined by the VLMC model can be written as

$$P_A \cdot P_{A|A} \cdot P_{A|G} \cdot P_{AG|T} \cdot P_{T|C} \cdot P_{TC|C} \cdot P_{CC|A} \cdot P_{A|A} \cdot P_{A|A}. \quad (6)$$

If the initial state AA is again considered fixed, the VLMC-based likelihood (6) becomes a corresponding expression as the earlier obtained likelihood (3) based on the partition of the enlarged state space.

In contrast to Example 2, in Example 1, the relevant part of the history will, for each state, involve the two preceding states. Consequently, the context mapping according to (4) will lead to a full second-order MC under the VLMC-based representation of sparsity, which illustrates that when the dependence structure of the Markov model does not allow a recursive factorization of the joint probability, a representation based on a partition of the transition probability distribution may lead to a considerably sparser model than a context-based representation. Reversely, a VLMC model is always representable as an SMC model, which is formally shown in the succeeding text together with a condition identifying when an SMC has an equivalent VLMC representation. A VLMC model is often represented in terms of a context tree, which specifies a pruning of the sample paths of a full MC of order m with respect to the set \mathcal{B} . The sample paths of a full MC of order m can be determined through the tree specified in the following definition.

Definition 3. Let $\{X_n\}_{n=0}^\infty$ be a time homogeneous MC of a finite order m . Denote by τ the tree of possible sample paths for $X_{n-m}, \dots, X_{n-1}, X_n = x_n$, where $X_n = x_n$ corresponds to the root node and the nodes at distance r from the root correspond to the J^r end points $X_{n-r} = x_{n-r}$ of each possible path $X_{n-r} = x_{n-r}, \dots, X_{n-1} = x_{n-1}$ to $X_n = x_n$.

Lemma 1. Let \mathcal{B}, \mathcal{P} be the set of contexts and transition probability distributions for a VLMC model of order m . The set \mathcal{B} defines a partition $S = (s_1, \dots, s_k)$ of \mathcal{X}^m such that the transition

probability vectors satisfy the equality $p_{i| \cdot} = p_{j| \cdot}$ for all pairs of states $\{i, j\} \in s_c, c = 1, \dots, k$. Further, any two VLMC models with distinct context sets \mathcal{B}_1 and \mathcal{B}_2 define distinct partitions S_1 and S_2 of \mathcal{X}^m .

Proof. By definition, all contexts in \mathcal{B} are distinct from each other, and a VLMC maps a subset $s_c \subset \mathcal{X}^m$ of the sample paths X_{n-m}, \dots, X_{n-1} in τ to a unique context B_c such that the equality (4) holds for all $p_{i| \cdot}$, where i is a state in \mathcal{X}^m and where the unique $b^{(r)}$ in B_c is a suffix of each $i \in s_c$. Because no sample path can map to two contexts in \mathcal{B} , a partition $S = (s_1, \dots, s_k)$ with k classes results from the mapping of all the states in \mathcal{X}^m , where each class has the stated property. The latter property can be established as follows. As \mathcal{B}_1 and \mathcal{B}_2 must differ at least with respect to a single element, let B_1 denote a context that is in \mathcal{B}_1 but not in \mathcal{B}_2 . Then, the class s_1 induced by the VLMC model with B_1 cannot exist in S_2 because no other suffix $X_{n-r} = x_{n-r}, \dots, X_{n-1} = x_{n-1}$ will correspond to the exactly same set of sample paths X_{n-m}, \dots, X_{n-1} . \square

Theorem 1. Let (S, \mathcal{P}) be an SMC. Then, there is an equivalent representation based on the set of contexts \mathcal{B} of a VLMC model if and only if there exists a unique context B_c with $b^{(r)}$, which is a suffix to all states i assigned to the same class s_c in S , for all $c = 1, \dots, k$.

Proof. Firstly, lemma 1 established that each VLMC model with a context set \mathcal{B} defines a unique partition S of the states in \mathcal{X}^m . Hence, by choosing the VLMC model that corresponds to the partition specified by the SMC, equality of the two representations follows by definition under the stated condition. To show that an SMC model lacks a VLMC representation if the stated condition is not satisfied, consider the tree τ of sample paths under the full MC of order m . Assume that a class contains two states $i = (x_{n-m}^{(i)}, \dots, x_{n-1}^{(i)})$ and $j = (x_{n-m}^{(j)}, \dots, x_{n-1}^{(j)})$ such that they do not share a suffix $b^{(r)}$ whose length is larger than zero. Then, there exists no context B_c , which maps the histories i and j such that they correspond to the same conditional probability $P(x_n | x_{n-1}^{(i)}, \dots, x_{n-m}^{(i)}) = P(x_n | x_{n-1}^{(j)}, \dots, x_{n-m}^{(j)})$. \square

Although some SMC models lack a context-based representation corresponding to a standard VLMC model, the definition of contexts can be generalized to accommodate dependence structures reflecting the class of SMC models, as stated in the following definition.

Definition 4. Generalized context set. Let S, \mathcal{P} be an SMC. Given (S, \mathcal{P}) , define the generalized context set $\mathcal{B} = \{B_c \subset \mathcal{X}^{r_c} : c = 1, \dots, k\}$ as the set of k contexts, such that for each class s_c in S , the suffixes $b_h^{(r)}$, $r \leq m$ of the strings in B_c indexed by $h = 1, \dots, |B_c|$ share a common prefix $a^{(l)}$, $l \leq r$ for which the constraint

$$P(x_n = j | b^{(r)}, x_{n-r-1}, \dots, x_{n-m}) = P(x_n = j | b^{(r)}) = P(x_n = j | a^{(l)}),$$

is satisfied. That is, the $r - l$ most recent states $x_{n-1}, \dots, x_{n-(r-l)}$ of the history are irrelevant for predicting x_n , whereas the l states $x_{n-(r-l)-1}, \dots, x_{n-r}$ need to be retained. Such a generalized context is denoted by $[A_{a^{(l)}}]a^{(l)}$, where $[A_{a^{(l)}}]$ denotes the set of states $x_{n-1}, \dots, x_{n-(r-l)}$ over which the conditional probabilities $P(x_n = j | b^{(r)})$ are identical.

Remark. In particular, when the cardinality of the generalized context set equals $|\mathcal{X}^{r-l}|$, the generalized context can be more compactly denoted as $[*]a^{(l)}$, because all possible paths from $a^{(l)}$ onwards lead to the same conditional probability for x_n . The generalized contexts enable

a compact representation of an MC that is sparse in the sense of an SMC. However, not all partitions of \mathcal{X}^m enjoy such an alternative representation, because a class of states s_c may lack entirely the possibility of finding a common prefix. For instance, $s_1 = \{AA, TT\}$ is an example of such a class for the DNA alphabet. Moreover, the generalized context representation does not necessarily lead to an identical likelihood as defined by the SMC, which is illustrated by the example in the succeeding text.

It is worth noticing that the concept of generalized contexts is not novel in itself. Wong & Ma (2010) derived a prior distribution that is a generalization of the Pólya tree approach by using optional stopping and optional choice of splitting variables. Another related recent construct is introduced in Dimitrakakis (2010b), who considers a generalization of contexts by a sequence of covers on the conditioning variable. Also, Roos & Yu (2009) mentioned in the case of a binary state space the possibility of using Haar transformation to specify models where a symbol has a zero effect on average, even when symbols further away in a context have non-zero effects, which has a clear connection to the sparse models presented here. However, they do not explicitly pursue the idea any further in experiments or theoretical considerations.

Example 4. As an illustration of the generalized context function, consider the model in Example 1 with the partition S :

$$S = \{\{AA, AC, AG, AT\}, \{CA, CC, CG, CT\}, \{GA, GC, GG, GT\}, \{TA, TC, TG, TT\}\}.$$

Notice that the ordering of the letters in strings within the cells of the aforementioned partition is reversed with respect to the ordering for generalized contexts. Here, the generalized context set equals $\mathcal{B} = \{[*]A, [*]C, [*]G, [*]T\}$ in the compact representation, and the corresponding likelihood expression for $(x_0x_1 \cdots x_8) = (AAGTCCAAA)$ becomes

$$P_{A*|G} \cdot P_{A*|T} \cdot P_{G*|C} \cdot P_{T*|C} \cdot P_{C*|A} \cdot P_{C*|A} \cdot P_{A*|A}, \quad (7)$$

when the probability of the initial state x_0x_1 is assumed fixed. Here, the two transitions from $x_5x_6 = CA$ to $x_7 = A$ and $x_4x_5 = CC$ to $x_6 = A$ have identical likelihoods with respect to the generalized context set. However, in the SMC-based likelihood given earlier, these transitions have the probabilities $\theta_{CA}^{(2)} = P_{CA|CA}$ and $\theta_{AA}^{(2)} = P_{CA|AA}$, which are not restricted to be equal, because the target state differs in the two cases.

3. Inference for sparse Markov chain models

Bayesian inference for VLMC models in terms of exact expressions for an unnormalized posterior has been earlier considered to a very limited extent. Corander *et al.* (2009) derived analytically the marginal likelihood of a data sequence with respect to a VLMC model and applied it to learning of the background noise model for *de novo* simultaneous detection of multiple classes of DNA regulatory binding regions. Dimitrakakis (2010a) considered Bayesian estimation of VLMC models using a method similar to the context tree weighting, and Bacallado (2011) introduced a conjugate prior for reversible VLMC models. Here, we derive first both prior predictive and posterior predictive distributions for SMC models, and then introduce a stochastic optimization algorithm to identify an approximate maximum *a posteriori* (MAP) model from data.

Consider an SMC model of order m defined by the pair (S, \mathcal{P}) , where we have k vectors of parameters $\{p_c\}$, $c = 1, \dots, k$. Let $\theta \in \Theta$ denote collectively the set of quantitative

parameters of an SMC model. Assuming the canonical conjugate multivariate Dirichlet prior for the matrix of transition probabilities (Koski, 2001), we have

$$p(\theta|\alpha, q) = \prod_{c=1}^k \left[\frac{\Gamma(\alpha)}{\prod_{j=1}^J \Gamma(\alpha q_j)} \prod_{j=1}^J p_{c|j}^{\alpha q_j - 1} \right], \quad (8)$$

where the hyperparameters satisfy the following conditions: $\alpha > 0$, $q_j > 0$, $\sum_{j=1}^J q_j = 1$. The likelihood of an observed data sequence $\mathbf{x} = x_0 x_1 \cdots x_n$ with initial state $z_0 = (x_0 x_1 \cdots x_{m-1})$ assumed fixed equals under the SMC model

$$p(\mathbf{x}|\theta, S) \propto \prod_{i=1}^{J^*} \prod_{j=1}^J p_{i|j}^{n_{i|j}} = \prod_{c=1}^k \prod_{j=1}^J p_{c|j}^{\sum_{i \in s_c} n_{i|j}}, \quad (9)$$

where $n_{i|j}$ is the observed count of transitions from the state i to j in \mathbf{x} . Consequently, the marginal likelihood $p(\mathbf{x}|S)$ of \mathbf{x} is available analytically using the properties of Dirichlet distribution, such that

$$\begin{aligned} p(\mathbf{x}|S) &\propto \int_{\theta \in \Theta} p(\mathbf{x}|\theta, S) p(\theta|\alpha, q) d\theta \\ &\propto \int_{\theta \in \Theta} \left[\prod_{c=1}^k \frac{\Gamma(\alpha)}{\prod_{j=1}^J \Gamma(\alpha q_j)} \prod_{j=1}^J p_{c|j}^{\alpha q_j - 1} \prod_{j=1}^J p_{c|j}^{\sum_{i \in s_c} n_{i|j}} \right] d\theta \\ &\propto \prod_{c=1}^k \frac{\Gamma(\alpha)}{\prod_{j=1}^J \Gamma(\alpha q_j)} \frac{\prod_{j=1}^J \Gamma(\sum_{i \in s_c} n_{i|j} + \alpha q_j)}{\Gamma(\left(\sum_{j=1}^J \sum_{i \in s_c} n_{i|j}\right) + \alpha)}, \end{aligned} \quad (10)$$

where $\Gamma(\cdot)$ is the Gamma function.

Conditional on m , posterior probability of S is obtained by assigning a prior distribution over the space of possible partitions of \mathcal{X}^m . For simplicity of implementation, we have used the uniform prior over the partition space, that is, $p(S) = 1/B_{|\mathcal{X}^m|}$, where $B_{|\mathcal{X}^m|}$ is the $|\mathcal{X}^m|$ th Bell number, in all the numerical experiments reported in the succeeding text. Similarly, the order parameter m is also assigned a uniform distribution over the values $m = 0, \dots, M$, where M is an upper bound, preferably specified using knowledge about reasonable values of the order in a particular application context. The joint posterior distribution of m and S is then defined by

$$p(S, m|\mathbf{x}) \propto p(\mathbf{x}|S) p(S) p(m), \quad (11)$$

where $p(m) = 1/(M + 1)$. Despite that a uniform prior distribution on S assigns more probability mass on a larger number of k than on small values because there are more partitions in the former case, the joint prior will still penalize an increase in the order m considerably because the probability mass $1/(M + 1)$ of any single value of order will be split evenly over an increasing set of partitions when m increases. If an inadequately small value of M would be chosen, the posterior distribution of m is likely to be concentrated at the upper bound, which provides a clear signal to re-consider the analysis on the basis of a larger M .

Using the predictive approach, we also find it possible to derive an analytical posterior predictive distribution for the future sequence of states $X_{n+1}, X_{n+2}, \dots, X_{n+l}$, conditional on S

and any particular sequence of previous states of the process $X_{n-m-1} = x_{n-m-1}, \dots, X_n = x_n$. The future sequence has the predictive probability

$$\begin{aligned} p(X_{n+l} = x_{n+l}, \dots, X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_{n-m-1} = x_{n-m-1}, S) &= \\ &= \int_{\theta \in \Theta} p(x_{n+l}, \dots, x_{n+1} | x_n, \dots, x_{n-m-1}, \theta, S) p(\theta | \mathbf{x}, S) d\theta \\ &= \prod_{c=1}^k \frac{\Gamma(\alpha + \sum_{i \in s_c} n_{i|j})}{\Gamma((\sum_{j=1}^J \sum_{i \in s_c} m_{i|j} + n_{i|j}) + \alpha)} \prod_{j=1}^J \frac{\Gamma(\sum_{i \in s_c} m_{i|j} + n_{i|j} + \alpha q_j)}{\Gamma(\sum_{i \in s_c} m_{i|j} + \alpha q_j)}, \end{aligned} \quad (12)$$

where $m_{i|j}$ is defined analogously to the count $n_{i|j}$ and is calculated from the sequence $\mathbf{x}_{n+l \dots n+1}$. The form of this probability follows from the conjugacy property of the multivariate Dirichlet distribution (Koski, 2001).

The MAP estimator of S conditional on m equals

$$\hat{S} = \arg \max_{S \in \mathcal{S}} p(\mathbf{x} | S) p(S), \quad (13)$$

which cannot in general be obtained using complete enumeration over the space of partitions due to its rapidly growing size. Assuming that an algorithm is available for calculating an approximate MAP estimate for a given m , then a MAP estimate over the class of SMC models is obtained by sequentially considering each value of order with positive prior probability:

$$(\hat{S}, \hat{m}) = \arg \max_{m \in \{0, \dots, M\}} \left\{ \arg \max_{S_m \in \mathcal{S}_m} p(\mathbf{x} | S_m) p(S_m) \right\}, \quad (14)$$

where $p(\mathbf{x} | S_m)$ and $p(S_m)$ denote the marginal likelihood and prior probability for a particular value of m , respectively, and \mathcal{S}_m refers to the space of possible partitions for a given order m . To ensure compatibility of model comparisons between all putative values of m , the initial observations $x_0 x_1 \dots x_{M-1}$ are considered fixed similar to order comparison for ordinary MC models.

Because the sufficient statistics from the data sequence \mathbf{x} correspond to an array of observed transition vectors defined in (9), the MAP partition estimation problem corresponds to clustering the transition data from each of the observed states in \mathcal{X}^m . In likelihood terms, this problem is identical to the population genetic problem of clustering populations (Corander *et al.*, 2003; Corander and Marttinen, 2006), when the number of available molecular marker loci equals one and the set of alleles in each population is defined as \mathcal{X} . Bayesian MC Monte Carlo (MCMC)-based clustering methods typically employing random split/merge/move-type search operators (e.g. Dawson and Belkhir, 2001; Corander *et al.*, 2003; Saraiva and Milan, 2012) represent a class of possible approaches to traverse the space of partitions to obtain the MAP estimate. However, such operators easily become numerically inefficient for larger state spaces, requiring impractically long simulations to be pursued, and therefore, one can alternatively use more efficient search operators that make intelligent data-driven proposals as discussed in, for example, Tu & Zhu (2002), Corander & Marttinen (2006) and Corander *et al.* (2006, 2008). Marttinen *et al.* (2006) and Marttinen *et al.* (2009) demonstrated that stochastic greedy optimization with such proposals could easily outperform standard MCMC approach to identify a competitive MAP estimate for challenging model classes.

Here, we adapt the stochastic greedy search algorithm considered in Corander & Marttinen (2006) and Marttinen *et al.* (2006) to the current learning problem. The adapted algorithm is based on the following steps for a given value of m :

- (i) initialize $S_t, t = 0$, with $|\mathcal{X}^m|$ singleton clusters and store for all pairs of states $i, l \in \mathcal{X}^m$ the distances between posterior mean estimates of their transition probability vectors

$$d_{i,l} = \sum_{j=1}^J \left(\frac{n_{i|j} + \alpha q_j}{\sum_{j=1}^J n_{i|j} + \alpha q_j} - \frac{n_{l|j} + \alpha q_j}{\sum_{j=1}^J n_{l|j} + \alpha q_j} \right)^2; \quad (15)$$

- (ii) given the current value of $p(\mathbf{x}|S_t)$, apply the following operators sequentially until no change in S_t results in a higher marginal likelihood;
- (iii) in a random order, move each state $i \in \mathcal{X}^m$ to the class c in S_t , which results in the S_{t+1} associated with a maximal increase in $p(\mathbf{x}|S_{t+1})$. If $p(\mathbf{x}|S_{t+1}) \leq p(\mathbf{x}|S_t)$ for all $c = 1, \dots, k$, $S_{t+1} = S_t$;
- (iv) for each pair of classes $c, c' = 1, \dots, k$, calculate $p(\mathbf{x}|S^*)$ for the S^* obtained by merging classes c, c' in S_t . If any S^* satisfies $p(\mathbf{x}|S^*) - p(\mathbf{x}|S_t) > 0$, set S_{t+1} equal to the S^* for which $p(\mathbf{x}|S^*) - p(\mathbf{x}|S_t)$ is maximal, otherwise set $S_{t+1} = S_t$;
- (v) for each class $c = 1, \dots, k$, use the complete linkage algorithm (e.g. Mardia *et al.*, 1979) with distances (15) to split the class into two non-empty subsets of states and calculate $p(\mathbf{x}|S^*)$ for the resulting partition S^* . If $p(\mathbf{x}|S^*) - p(\mathbf{x}|S_t) > 0$, set S_{t+1} equal to S^* , otherwise set $S_{t+1} = S_t$.

Because the distances between the posterior mean estimates of transition probability vectors can be calculated and stored before the stochastic search, the split operator can make intelligent data-driven splits of clusters in contrast to a standard MCMC-type operator proposing random splits. The stochastic search algorithm converges to a local mode when no improvements to the posterior probability are accessible under the given operators. To increase the probability of finding globally representative areas of the posterior, the algorithm is used with restarts from different initial configurations having less than $|\mathcal{X}^m|$ clusters. To facilitate escape from eventual local peaks of the posterior that cannot be overcome by the aforementioned operators, it would also be possible to generalize this algorithm into a hybrid version similarly to Marttinen *et al.* (2006), where stochastic greedy optimization was interleaved with non-reversible MCs. This strategy would yield a consistent posterior mode estimator given the results from Corander *et al.* (2006, 2008), but we do not pursue it further here because of the added computational complexity and the already satisfactory results obtained using the current version.

To illustrate the predictive ability of SMC models, we applied the greedy stochastic learning algorithm to both synthetic and real DNA sequences with $\mathcal{X} = \{A, C, G, T\}$. This is a relevant area of application because techniques for tasks such as sequence segmentation and classification benefit from efficient modelling of transition probabilities in the DNA. Further, advances in these tasks open new possibilities for biological and medical research. SMC models seem to be less suitable for some other domains of application. We noticed, for example, that processing of natural language, such as, ASCII text (with the alphabet size of 128), the computer memory needed for an implementation of an SMC learning algorithm is impractically high even for models of moderate order. Besides DNA, promising and computationally feasible applications include, for example, classification of protein sequences, which will be explored at the end of this section.

For empirical assessment of SMC, we compared it with other VOM models for data compression for which implementations were publicly available and which can handle also larger than binary state spaces. These were the Lempel-Ziv 78 method and a recent modification of it (LZ-MS), prediction by partial match method-C (PPMC), decomposed context tree weighting (DCTW) and binary context tree weighting as described and reviewed in Begleiter *et al.* (2004)

where also a reference to Java/Matlab implementation of the algorithms is given. PPMC and DCTW were reported as the generally strongest algorithms in the original article, and therefore, our comparisons focus on them. LZ-MS is not included here because its performance in prediction as reported in Begleiter *et al.* (2004) was mediocre. However, in the classification task, LZ-MS was reported to achieve very good results. So for classification, we include LZ-MS as one of the algorithms in the comparison.

We combined our implementation of the SMC learning algorithm with the existing implementations of the other algorithms to get a train/test protocol for DNA sequence data. In the training phase, each algorithm was trained with the training sequence, whereas the model parameters were tuned with fivefold repetition similar manner to what was described in Begleiter *et al.* (2004). This was performed to get more representative view of the capabilities of the algorithms. In the testing phase, average log-loss was calculated over the testing sequence. This is a way to measure the predictive performance of the estimated probability model. For the test sequence $x = x_1 \dots x_T$, the average log-loss is

$$l(\hat{P}, x) = -\frac{1}{T} \sum_{i=1}^T \log_2 \hat{P}(x_i | x_1 \dots x_{i-1}), \quad (16)$$

where the conditional probabilities of the form $\hat{P}(x_i | x_1 \dots x_{i-1})$ are given by the model estimated from training data (Begleiter *et al.*, 2004). This calculation of the average log-loss is possible for all the models considered here.

To produce synthetic DNA data under an ordinary MC with order $m = 3$, parameters for the generating model were sampled as follows. Firstly, a four-dimensional multivariate normal distribution with zero mean vector and the covariance matrix were defined as $5 \cdot I_4$ (I_4 refers to a 4×4 identity matrix). This was used for generating vectors $\mathbf{z}_l = (z_{l1}, z_{l2}, z_{l3}, z_{l4})$ with $l = 1, \dots, 64$. Distributions of the transitions probabilities were then sampled from the corresponding Dirichlet ($e^{z_{l1}}, e^{z_{l2}}, e^{z_{l3}}, e^{z_{l4}}$) distributions for each of the 64 states. To simulate parameters for an SMC model with order $m = 3$, 20 transition probability vectors were generated in an analogous manner to the previous case. These were assigned with uniform probabilities to the 64 states of the third-order MC. Under both generating models, 200 independent MCs of length 10^5 were simulated, and the log-losses were calculated for each realization sequentially using an increasing amount of training data. In Fig. 1, log-loss distributions are summarized for the competing algorithms when the generating model is a third-order MC. Relative log-loss has been calculated by subtracting the average log-loss score of the generating model from that of the learned model. Despite of the fact that the generating model is not sparse, our SMC-based method performs clearly best. The difference in predictive ability in favour of SMC becomes even more pronounced when the underlying model is sparse, which is illustrated in Fig. 2.

To investigate how well our SMC method performs in terms of parameter estimation, we generated data using the model in Example 1 in the previous section. In that model, there are four classes of transition probability vectors and $m = 2$. The model parameters were generated using otherwise the same aforementioned procedure, but with correspondingly reduced size of the state space ($|\mathcal{X}^m| = 16$). We applied the SMC estimation method to 50 simulated sequences of length 5000 observations, extracted the posterior means of the transition probability distributions for all classes and calculated the minimum mean squared error over all possible permutations of the resulting partition classes when $k = 4$, because the classes are unordered, and it cannot be known directly which inferred class corresponds to which class in the generating model. The MAP estimate had $k = 4$ in 45/50 cases, and the remaining data sequences did lead to $k = 5$ as the optimal estimate. Figure 3 illustrates the decrease of mean

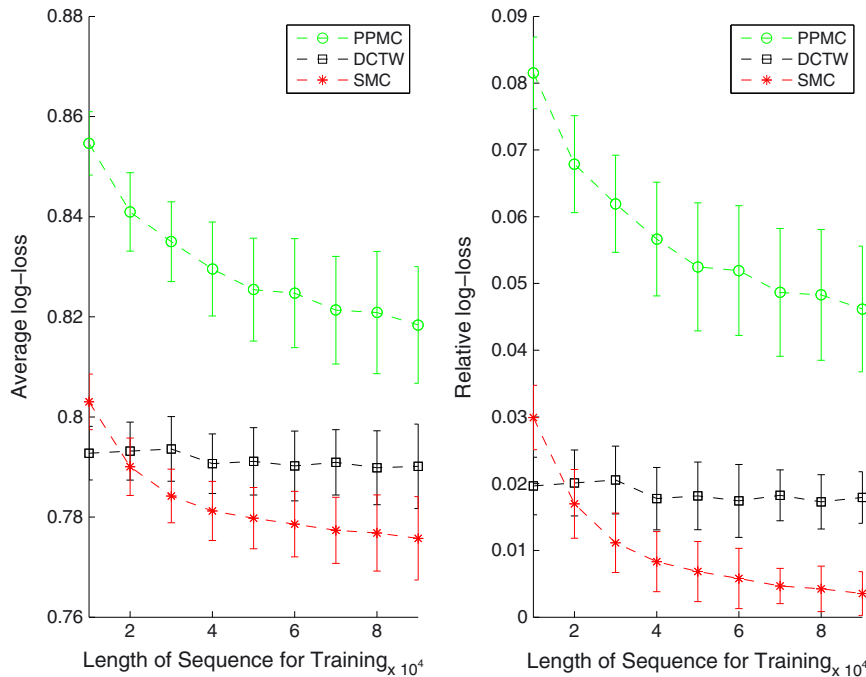


Fig. 1. Means \pm standard deviations (interval endpoints) for log-loss over 200 replicates of a Markov chain of order $m = 3$.

squared error of the parameter estimates as a function of the amount of data available for the inference method.

For studying real DNA data, we used a large bacterial genomic database investigated in detail in Corander *et al.* (2012), where each sequence was assigned to a cluster using a population genomic model. Two experiments were performed using the concatenated multilocus sequence typing (MLST) DNA sequences of 7829 *Neisseria meningitidis* bacterial strains. In the first example, 200 sets of sequences were sampled randomly with replacement from the database. Each set included two sequences for training and two for testing. Within a single set, the four sequences were all from different clusters identified in Corander *et al.* (2012). In addition to that, all the sequences used in the experiment were unique in terms of their nucleotide data. Because the generating model in this case is not known, comparison between algorithms is more challenging than in the previous examples. We used a training sequence with a cumulatively increasing length while keeping the testing data constant. Thus, within a single set, the two test sequences were used for all lengths of the training sequence. This procedure was repeated separately for the 200 sets, and log-loss was calculated after each interval obtained by segmentation of the sequences into 20 equally sized parts. This experiment is particularly challenging because the amount of training data is very small. For the first half of the sequences, the three best methods provide indistinguishable results, whereafter DCTW improves gradually over SMC and PPMC.

To investigate how an increase in the amount of training data affects the performance of the different methods, the data were randomly split into 3915 training and 3914 test sequences. All the same algorithms as reported in Fig. 4 were again used; however, we also added a full MC model and considered orders between 5 and 10. Because of the extensive computation times, the experiment was only performed once. Table 1 displays the log-losses only for ordinary MCs

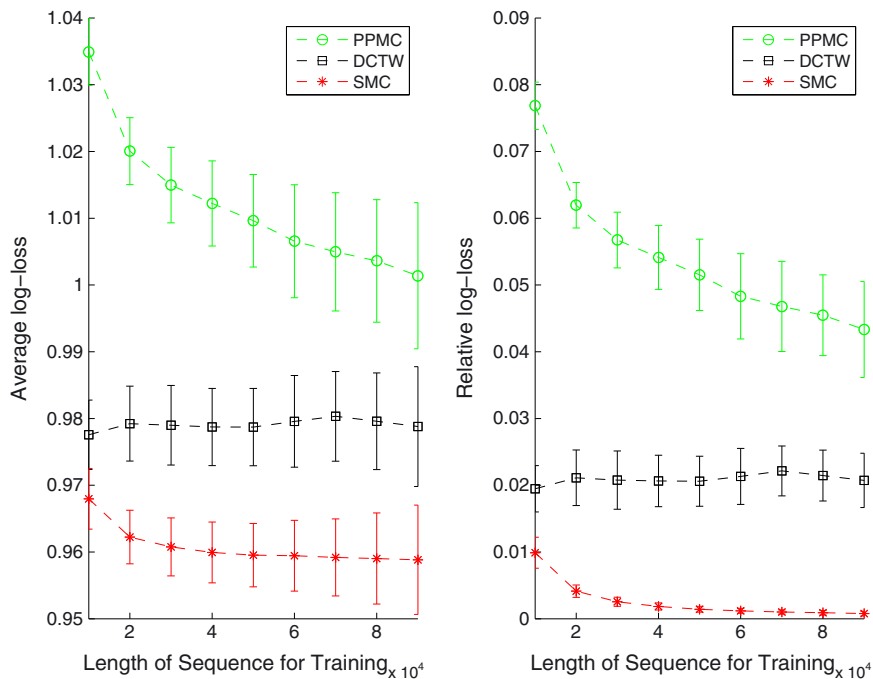


Fig. 2. Means \pm standard deviations (interval endpoints) for log-loss over 200 replicates of a sparse Markov chain with order $m = 3$ and 20 classes of transition probability distributions.

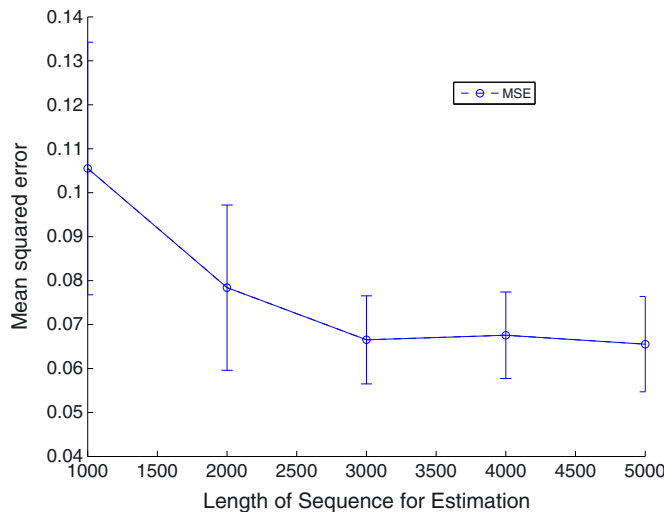


Fig. 3. Means \pm standard deviations (interval endpoints) for mean squared error of parameter estimates for a Markov chain of order $m = 2$.

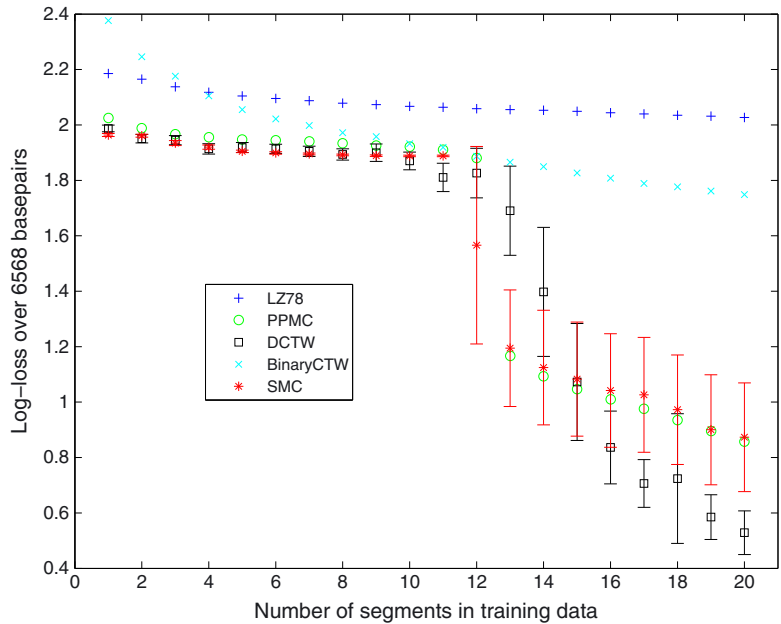


Fig. 4. Means (\pm standard deviations shown for decomposed context tree weighting and sparse Markov channel) of log-loss over 200 replicates of the real DNA data.

Table 1. Log-loss for predicting the 3914 concatenated multilocus sequence typing DNA sequences based on 3915 training sequences

Order	DCTW	MC	PPMC	SMC
5	1.424	1.629	1.403	1.374
6	0.928	1.542	0.930	0.882
7	0.483	1.505	0.492	0.448
8	0.228	1.504	0.246	0.209
9	0.119	1.513	0.128	0.108
10	0.080	1.526	0.090	0.073

DCTW, decomposed context tree weighting; MC, Markov chain; PPMC, prediction by partial match method-C; SMC, sparse Markov chain.

and for DCTW, PPMC and SMC, as the other methods had so low scores that they are excluded from this comparison. As expected, the sparse methods are superior to an ordinary MC for all orders. SMC shows here consistently the best predictive performance, similar to the simulated DNA data experiments.

Using log-loss for measuring model performance has its limitations for certain applications. For example, Begleiter *et al.* (2004) reported that average log-losses in the protein prediction task were larger than entropy of the uniform distribution. When they classified protein sequences instead of predicting, LZ-MS achieved the best results. But when in an additional setup, prediction was based on classification, DCTW and PPMC dominated. This variability suggests that several testing procedures give better understanding of the algorithms than prediction alone. We tested our approach of SMCs with the classification of the protein sequences.

Table 2. Error rates in protein classification. Values for the LZ-MS and decomposed context tree weighting; methods are taken from Begleiter *et al.* (2004)

Class	LZ-MS		DCTW		SMC	
	Mean	Std	Mean	Std	Mean	Std
A	0.16	0.031	0.19	0.031	0.21	0.018
B	0.14	0.031	0.17	0.031	0.16	0.01
C	0.17	0.035	0.22	0.031	0.18	0.02
D	0.16	0.031	0.19	0.031	0.17	0.08
E	0.05	0.017	0.09	0.031	0.04	0.025
F	0.17	0.03	0.21	0.044	0.14	0.044
G	0.16	0.017	0.25	0.035	0.21	0.053

DCTW, decomposed context tree weighting; SMC, sparse Markov chain; Std, standard deviation.

The test protocol followed that of Begleiter *et al.* (2004) so that a direct comparison of the results is possible.

Table 2 presents error rate (1-accuracy) with standard deviations. In the original article, LZ-MS was reported as the most successful algorithm, and DCTW was also among the better ones. In Table 2, LZ-MS has the best average performance in test classes A, B, C, D and G, whereas SMC gives the best accuracy in classes E and F. In all except one of the test classes, SMC achieves error rates that are lower than those of DCTW. Although LZ-MS has here slightly smaller average error rate than SMC, its performance for the DNA sequence prediction was inferior, and the results generally illustrate that SMC can achieve performance that is at a comparable level with the best algorithms reviewed in Begleiter *et al.* (2004) and even markedly better under certain circumstances.

4. Discussion

Versatility of VOM and VLMC models has been demonstrated in a multitude of applications to data compression and prediction, ranging from text modelling to describing background variation in DNA sequences. Very recently, a considerable interest has surfaced to generalize this class of models and to introduce novel learning methods for the new models. Using results derived earlier for VLMC models, we here introduced a way to perform Bayesian inference for SMC models, which lump transition probabilities into invariant classes, such that the resulting models need not have a hierarchical structure as in general required in context tree-based approaches. We illustrated with synthetic and real DNA data, as well as by classification of protein sequences, that our approach leads to good performance when compared with earlier proposed algorithms. None of the algorithms considered here showed clearly superior performance to all other methods throughout the testing, which emphasizes the challenging nature of the DNA and protein data types.

In our approach, we used a default uniform prior for the partitions of sequence states, combined with a uniform prior for the MC order. The experiments indicate that this choice leads to satisfactory results, because the prior for transition probability parameters penalizes too excessive models. However, more advanced priors directly penalizing an increase in the number of classes could also be considered, for instance, using a Dirichlet process model (Neal, 2000).

DNA sequence data offer an attractive target of application for the SMC models, given that higher order MCs will typically be too parameter rich for reliable estimation. More sparse VOM/VLMC models have already earlier been demonstrated to offer clear benefits over MCs in various advanced bioinformatics applications (Ben-Gal *et al.*, 2005; Zhang *et al.*, 2005; Browning, 2006; Corander *et al.*, 2009). It would thus be interesting in the future to generalize

the SMC models to accommodate various types of situations where SMCs represent modular parts of the total model, such as model-based sequence segmentation, clustering of sequences and regulatory element identification.

Acknowledgements

The authors would like to thank the two anonymous reviewers whose comments and suggestions did lead to significant improvements over an earlier version of this article.

Work of JC, JX and VJ was funded by the ERC grant no. 239784 and by the Academy of Finland grant no. 251170; work of VJ and JX was also funded by the FICS and FDPSS graduate schools.

References

- Bacallado, S. (2011). Bayesian analysis of variable-order, reversible Markov chains. *Ann. Stat.* **39**, 838–864.
- Begleiter, R., El-Yaniv, R. & Yona, G. (2004). On prediction using variable order Markov models. *J. Artif. Intell. Res.* **22**, 385–421.
- Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S., Grosse, I. & Journals, O. (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics* **21**, 2657–2666.
- Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* **78**, 903–913.
- Bühlmann, P. & Wyner, A. J. (1999). Variable length Markov chains. *Ann. Stat.* **27**, 480–513.
- Corander, J., Waldmann, P., Marttinen, P. & Sillanpää, M. J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**, 367–374.
- Corander, J. & Marttinen, P. (2006). Bayesian identification of admixture events using multi-locus molecular markers. *Mol. Ecol.* **15**, 2833–2843.
- Corander, J., Gyllenberg, M. & Koski, T. (2006). Bayesian model learning based on a parallel MCMC strategy. *Stat. Comput.* **16**, 355–362.
- Corander, J., Ekdahl, M. & Koski, T. (2008). Parallell interacting MCMC for learning of topologies of graphical models. *Data Min. Knowl. Disc.* **17**, 431–456.
- Corander, J., Ekdahl, M. & Koski, T. (2009). Bayesian unsupervised learning of DNA regulatory binding regions. *Adv. Art. Int.* Vol. 2009, 11, Article ID 219743. DOI:10.1155/2009/219743.
- Corander, J., Connor, T. R., O'Dwyner, C. A., Kroll, J. S. & Hanake, W. P. (2012). Population structure in the Neisseria, and the biological significance of fuzzy species. *J. R. Soc. Interface* **9**, (71), 1208–1215.
- Csiszár, I. & Shields, P. C. (2000). The consistency of the BIC Markov order estimator. *Ann. Stat.* **28**, 1601–1619.
- Csiszár, I. & Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes via BIC and MDL. *IEEE Trans. Inf. Theory* **52**, 1007–1016.
- Dawson, K. J. & Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet. Res., Camb.* **78**, 59–77.
- Dimitrakakis, C. (2010a). Bayesian variable order Markov models. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, **9**, 161–168, Chia Laguna Resort, Sardinia, Italy.
- Dimitrakakis, C. (2010b). Context models on sequences of covers. arXiv preprint arXiv:1005.2263.
- Farcomeni, A. (2011). Hidden Markov partition models. *Stat. Probabil. Lett.* **81**, 1766–1770.
- Gasthaus, J., Wood, F. & Teh, Y. W. (2010). Lossless compression based on the sequence memoizer. In *Data Compression Conference*, 337–345, Utah, USA.
- García, J. E. & González-López, J. E. (2010). Minimal Markov models. arXiv preprint arXiv:1002.0729.
- Koski, T. (2001). *Hidden Markov models for bioinformatics*, Kluwer, Dordrecht.
- MacKay, D. J. C. & Peto, L. (1995). A hierarchical Dirichlet language model. *Nat. Lang. Eng.* **1**, 1–19.
- Mardia, K. V., Kent, J. T. & Bibby, J. M. (1979). *Multivariate analysis*, Academic Press, London.
- Marttinen, P., Corander, J., Törönen, P. & Holm, L. (2006). Bayesian search of functionally divergent protein subgroups and their function specific residues. *Bioinformatics* **22**, (20), 2466–2474.

- Marttinen, P., Myllykangas, S. & Corander, J. (2009). Bayesian clustering and feature selection for cancer tissue samples. *BMC Bioinformatics* **10**, 90.
- Mächler, M. & Bühlmann, P. (2004). Variable length Markov chains: methodology, computing, and software. *J. Comput. Graph. Stat.* **13**, 435–455.
- Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comp. Graph. Stat.* **9**, 249–265.
- Rissanen, J. (1983). A universal data compression system. *IEEE Trans. Inf. Theory* **29**, 656–664.
- Roos, T. & Yu, B. (2009). Sparse Markov source estimation via transformed Lasso. In *Proceedings of the IEEE Information Theory Workshop (ITW-2009)*, 241–245, Taormina, Sicily, Italy.
- Saraiva, E. F. & Milan, L. A. (2012). Clustering gene expression data using a posterior split-merge-birth procedure. *Scand. J. Stat.* **39**, 399–415.
- Tu, Z. & Zhu, S. C. (2002). Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 657–673.
- Weinberger, M. J., Rissanen, J. & Feder, M. Ž. (1995). A universal finite memory source. *IEEE Trans. Inf. Theory* **41**, 643–652.
- Willems, F. M. J., Shtarkov, Y. M. & Tjalkens, T. J. (1995). The context tree weighting method: basic properties. *IEEE Trans. Inf. Theory* **41**, 653–664.
- Wong, W. H. & Ma, L. (2010). Optional Pólya tree and Bayesian inference. *The Ann. Stat.* **38**, 1433–1459.
- Wood, F., Archambeau, C., Gasthaus, J., James, L. & Teh, Y. W. (2009). A stochastic memoizer for sequence data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1129–1136, Montreal, QC, Canada.
- Zhang, X., Huang, H., Li, M. & Speed, T. (2005). Finding short DNA motifs using permuted Markov models. *Bioinformatics* **21**, 894–906.

Received August 2012, in final form September 2013

Jukka Corander, Department of Mathematics, Åbo Akademi University, FIN-20500, Finland.
E-mail: jukka.corander@abo.fi