1416–1424 Nucleic Acids Research, 2013, Vol. 41, No. 3 Published online 24 December 2012
doi:10.1093/nar/gks1285

One size does not fit all: On how Markov model order dictates performance of genomic sequence analyses

Leelavati Narlikar[1,2], Nidhi Mehta[1], Sanjeev Galande[3,4,*] and Mihir Arjunwadkar[1,5,*]

[1]Centre for Modeling and Simulation, University of Pune, Pune 411 007, [2]Chemical Engineering and Process Development Division, CSIR-National Chemical Laboratory, Pune 411 008, [3]National Centre for Cell Science, University of Pune Campus, Pune 411 007, [4]Indian Institute of Science Education and Research, Pune 411 021 and [5]National Centre for Radio Astrophysics, TIFR, University of Pune Campus, Pune 411 007, India

Received September 19, 2011; Revised October 24, 2012; Accepted November 10, 2012

ABSTRACT

The structural simplicity and ability to capture serial correlations make Markov models a popular modeling choice in several genomic analyses, such as identification of motifs, genes and regulatory elements. A critical, yet relatively unexplored, issue is the determination of the order of the Markov model. Most biological applications use a predetermined order for all data sets indiscriminately. Here, we show the vast variation in the performance of such applications with the order. To identify the 'optimal' order, we investigated two model selection criteria: Akaike information criterion and Bayesian information criterion (BIC). The BIC optimal order delivers the best performance for mammalian phylogeny reconstruction and motif discovery. Importantly, this order is different from orders typically used by many tools, suggesting that a simple additional step determining this order can significantly improve results. Further, we describe a novel classification approach based on BIC optimal Markov models to predict functionality of tissue-specific promoters. Our classifier discriminates between promoters active across 12 different tissues with remarkable accuracy, yielding 3 times the precision expected by chance. Application to the metagenomics problem of identifying the taxum from a short DNA fragment yields accuracies at least as high as the more complex mainstream methodologies, while retaining conceptual and computational simplicity.

INTRODUCTION

Genomes are complex hierarchically organized entities shaped largely through the forces of evolution. As a result, the primary sequence of a genome contains both short- and long-range correlations (1). Short-range correlations on the scale of a few base pairs are usually associated with the machinery for gene expression and its control (2,3), whereas long-range correlations are typically related to the properties of chromatin organization (4,5). Furthermore, functionally distinct genomic elements, such as promoters, introns, exons, intergenic regions, repetitive elements and regulatory elements are known to possess distinct sequence features (6,7). Moreover, genomes that are well separated during evolution have unique statistical characteristics of their own (8). These distinct statistical properties of genomes are often studied using probabilistic models.

A simple probabilistic model of a genomic sequence assumes independent, identically distributed genomic alphabet {A,C,G,T} occurring with specific probabilities. This model, however, does not account for correlations within a sequence. A more general model that captures sequential correlations in a systematic manner is a Markov model. In the genomic context, Markov models are specified in the form of a matrix of conditional probabilities connecting the set of all length-k genomic words to the genomic alphabet. In other words, under a Markov model, the probability of observing a particular nucleotide at a given position along a sequence depends only on the previous k nucleotides. This prespecified word length k is referred to as the *order* of the Markov model. If this probability also depends on the position within the sequence, the model is called an *inhomogeneous* Markov

*To whom correspondence should be addressed. Tel: +91 20 25719445; Fax: +91 20 25692149; Email: mihir@cms.unipune.ac.in
Correspondence may also be addressed to Sanjeev Galande. Tel: +91 20 25908060; Fax: +91 20 25865315; Email: sanjeev@iiserpune.ac.in

model. In this work, we focus on *homogeneous* Markov models, where all positions in the sequence are described by the same set of conditional probabilities.

Markov models have been used extensively in a variety of genomic sequence analysis contexts, such as probabilistic motif discovery (9), prediction of CpG islands through discrimination (10), computational gene finding (11), searching for RNA structures (12), sequence similarity measures (13,14), alignment-free sequence comparison (15) and genome segmentation (16).

A critical statistical issue, often overlooked in biological sequence analysis contexts, is the determination of an optimal value for the order k of the Markov model. While a higher order model allows greater complexity to be captured in the model, too high an order leads to overfitting. In contrast, simpler low-order models tend to be better determined from data, but too low an order may miss out on essential sequence features in the data. While in some cases, such as motif discovery, a higher order Markov model is shown to yield more accurate motifs (17), the question of which order will perform the best has not been addressed systematically thus far.

Here, we have treated the problem of identifying the optimal Markov order as a problem of selecting an optimal model to describe a given sequence data set. We used two well-known model selection criteria to tackle this problem; namely, the Akaike information criterion (AIC) (18) and the Bayesian information criterion (BIC), also known as the Schwarz criterion (19). In both approaches, each model in the set of models being considered is scored using the difference of two terms; namely,

(1) the likelihood of the data under the model, i.e. how well the model describes the training data and
(2) a monotonically increasing penalty on the number of model parameters, i.e. the complexity of the model.

The difference between AIC and BIC scores lies in the second term: while the AIC incorporates, as penalty, a simple linear function of the number of parameters, the BIC weighs the number of parameters by the size of the training data. The model with the minimum score is considered optimal under the criterion used. The optimal model therefore attempts to strike a balance between complexity and descriptive ability over the set of models considered, following the spirit of Occam's razor. The AIC-predicted optimal (APO) and BIC-predicted optimal (BPO) orders are obtained by evaluating a score (AIC or BIC) for all model orders under consideration. The order that minimizes a given score is considered optimal.

Considering the wide-spread use of Markov models in genomic sequence analysis, this naturally leads to the following key questions:

(1) how do different genomic sequence analysis methods behave with respect to the Markov order used?
(2) which model selection method leads to more biologically relevant results in typical genomic sequence analysis contexts? and
(3) how does the optimal order change with respect to the size of the sequence data?

In this study, we investigated the behavior, as a function of the model order, of the Markov model-based genomic sequence analysis methods in three broad biological contexts; namely, phylogeny reconstruction and metagenomics, probabilistic *de novo* motif discovery and functional classification of genomic sequences. In each case studied, we found that Markov models of the BPO order deliver the best performance. We argue that using the optimal order can make a significant difference to the results of genomic sequence analyses and, specifically, such order selection considerations can have serious repercussions in the context of metagenomics. We also show that a simple multiclass classifier incorporating the BPO order yields surprisingly accurate results for the challenging problem of distinguishing between promoters of tissue-specific genes, as well as for the binning problem of metagenomics, i.e. the identification of a prokaryote from a short DNA fragment. This demonstrates how simple homogeneous Markov models of genomic sequences, when built using the BPO order, can be remarkably informative.

MATERIALS AND METHODS

Building Markov models from DNA sequence data

Markov model $\mathcal{M}_k$ of order k representing a set of DNA sequences is represented in the form of a $4^k \times 4$ matrix of conditional probabilities $P(b|b_1b_2\ldots b_k)$ of a single base b to follow the length-k base sequence $b_1b_2\ldots b_k$. Here, b and $b_i, 1 \leq i \leq k$, take values from the DNA alphabet $\mathcal{A} = \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$. An order 0 Markov model thus corresponds to the aggregate probabilities of occurrence of the four DNA bases. By construction, elements of each row of the Markov matrix are non-negative and add up to 1.

Given a set $\mathcal{S}$ consisting of N genomic sequences $S_j \equiv b_1^{(j)}\ldots b_{L_j}^{(j)}$ of length L_j, $1 \leq j \leq N$, the log-likelihood for $\mathcal{S}$ under the model $\mathcal{M}_k$ takes the form

$$\lambda(\mathcal{S};\mathcal{M}_k) = \sum_{b,b_1,\ldots,b_k \in \mathcal{A}} n(b_1\ldots b_k b)\log P(b|b_1\ldots b_k)$$

where $n(b_1\ldots b_k b)$ is the total number of occurrences of the genomic word $b_1\ldots b_k b$ in the sequence set $\mathcal{S}$. There are a total of $4^k \times 3$ free parameters in such a model, which we estimate using the standard maximum likelihood approach. See Supplementary Methods for further details.

Markov model order selection using AIC/BIC

For a set of models $\widehat{\mathcal{M}}_k, 0 \leq k \leq k_{\max}$, estimated from the *same* sequence set $\mathcal{S}$, the AIC and the BIC are defined, respectively, as

$$\mathrm{AIC}(k) = -2\lambda(\mathcal{S};\widehat{\mathcal{M}}_k) + 2|\mathcal{M}_k| \quad \mathbf{(1)}$$

$$\mathrm{BIC}(k) = -2\lambda(\mathcal{S};\widehat{\mathcal{M}}_k) + |\mathcal{M}_k|\log|\mathcal{S}|_k, \quad \mathbf{(2)}$$

where $|\mathcal{S}|_k$ is the data size, i.e. the total number of length-$(k+1)$ words in $\mathcal{S}$. If a sequence in $\mathcal{S}$ contains the character N representing an undetermined base, we consider this sequence as being broken up into smaller fragments devoid of Ns. Optimal order k_* is found by minimizing

AIC(k) or BIC(k) with respect to $0 \le k \le k_{\max}$, where $k_{\max}$ is an arbitrary upper bound on the order k. see Supplementary Methods for further details.

2.3 Multiclass classification using Markov models

Consider a sequence set $\mathcal{S}$ composed of two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ assumed to be modeled with Markov models $\widehat{\mathcal{M}}_k^{(1)}$ and $\widehat{\mathcal{M}}_k^{(2)}$, respectively, of order k. Consider the problem of identifying the label l (1 or 2) of sequence S in $\mathcal{S}$. Durbin *et al.* (10) use the log odds criterion, which is equivalent to assigning the label as:

$$l(S) = \begin{cases} 1, & \text{if } \lambda(S; \widehat{\mathcal{M}}_k^{(1)}) > \lambda(S; \widehat{\mathcal{M}}_k^{(2)}) \\ 2, & \text{otherwise.} \end{cases}$$

In the multiple class situation, this amounts to predicting the label l $(= 1, \ldots, C)$ for sequence $S \in \mathcal{S}$, where the sequence set $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2 \ldots \cup \mathcal{S}_C$ consists of C classes of sequences. Assuming we have corresponding models $\widehat{\mathcal{M}}_k^{(1)}, \ldots, \widehat{\mathcal{M}}_k^{(C)}$, we generalize the above prescription to

$$l(S) = \arg\max_c \; \lambda(S; \widehat{\mathcal{M}}_k^{(c)}). \qquad \textbf{(3)}$$

This prescription is built on the intuitive notion of selecting a class label corresponding to the model under which the test sequence has the greatest probability of occurrence. Further details can be found in the Supplementary Methods. In principle, this approach can be extended to compare models of different orders, but this adds a layer of complexity to the formalism and the computation, which needs to be evaluated for its effectiveness. Therefore, in this article, we build our classifiers using the same order for all classes.

We evaluate the classifier using 5-fold cross-validation. In the metagenomics problem, we use only those clades that have genomic sequences for at least 10 organisms. The accuracy in the case of the promoter classification problem is computed as the total percentage of promoters predicted correctly. Since the number of promoters in each class (tissue) is the same, we do not need to normalize this quantity. In the metagenomics context, however, we use the class-normalized sensitivity (20) as a measure of accuracy, since the number of species in each clade is highly variable (see Supplementary Results).

RESULTS

APO and BPO orders for select eukaryotic genomes

To explore the behavior of the AIC- and BIC-based model selection methods, we computed the AIC and BIC scores for a selection of eukaryotic genomes. The behavior of these scores as functions of the model order is illustrated in Figure 1 for the fruitfly, chicken, zebrafish, opposum and human genomes. Both methods lead to trends showing a dip that identifies the optimal order. Table 1 further provides the APO and BPO orders for a number of additional genomes. For each genome explored here, the BPO order is strictly smaller than the APO order because of the larger penalty in the BIC. Generally, larger genomes tend to result in larger optimal orders under either criterion.

We also explored the utility of whole-genome Markov models for the purpose of reconstructing phylogeny via clustering. For illustrative purposes, we used 10 mammalian genomes across 4 different taxonomic orders. Detailed results of this exercise are included in Supplementary Results. Our reconstructed phylogeny matches the known biological classification for models constructed using orders 9 and above. Interestingly, the BPO order for all these genomes is 10 (Table 1), which

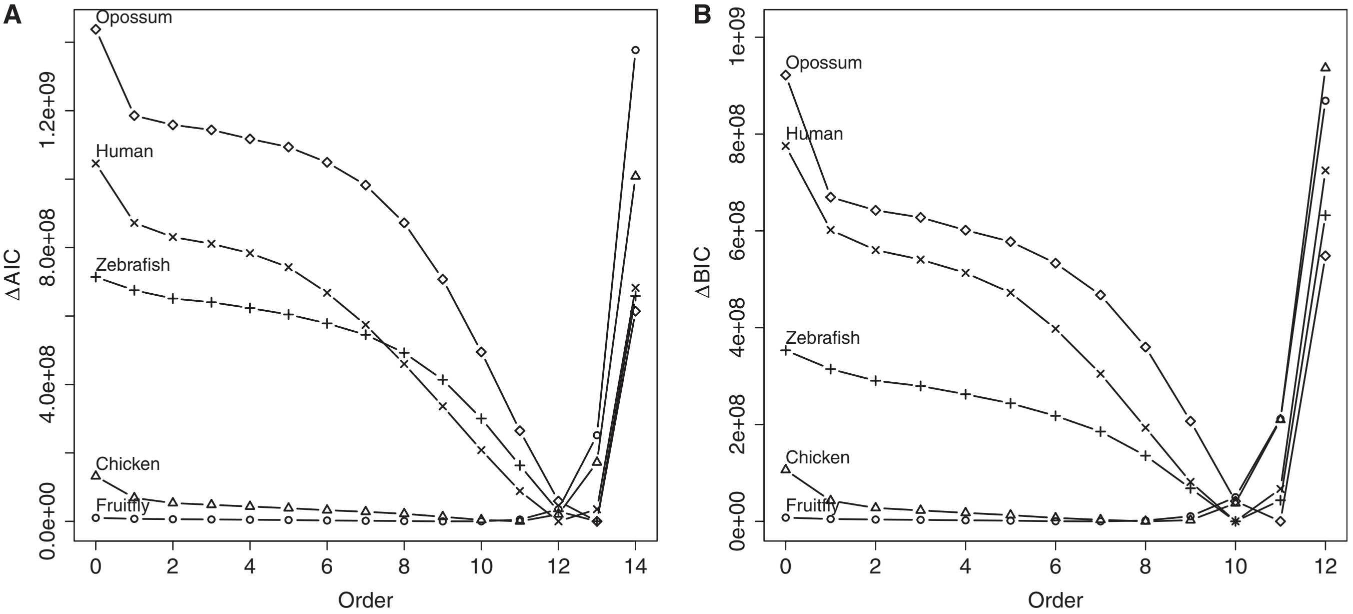

Figure 1. The AIC- and BIC-based order selection procedure illustrated. The *y*-axis represents the (**A**) AIC or (**B**) BIC score minus its minimum value for given genome. The order that minimizes a given score is considered optimal; this is the AIC- or the BIC-predicted optimal order. Optimal orders for these and other complete genomes are listed in Table 1.

Table 1. APO and BPO orders for a number of genomes, arranged by the increasing genome length

Genome (UCSC Version)	Length (Mb)	APO order	BPO order
Yeast (sacCer2)	12.2	7	4
Nematode (ce6)	100.3	10	7
Fruitfly (dm3)	120.3	9	7
Fugu (fr2)	351.2	11	8
Stickleback (gasAcu1)	446.6	11	8
Chicken (galGal3)	984.9	11	8
Zebra Finch (taeGut1)	1112.7	11	8
Xenopus (xenTro2)	1359.4	12	10
Zebrafish (danRer5)	1523.3	13	10
Cat (felCat3)	1642.7	12	10
Lizard (anoCar1)	1741.5	12	10
Platypus (ornAna1)	1842.2	12	10
Dog (canFam2)	2385.0	12	10
Horse (equCab2)	2428.8	12	10
Rat (rn4)	2533.3	12	10
Mouse (mm9)	2558.5	12	10
Macaque (rheMac2)	2646.7	12	10
Guinea pig (cavPor3)	2663.4	13	10
Cow (bosTau4)	2731.8	12	10
Orangutan (ponAbe2)	2788.0	12	10
Chimp (panTro2)	2802.8	12	10
Human (hg18)	2858.0	12	10
Marmoset (calJac1)	2929.1	12	10
Opossum (monDom5)	3501.7	13	11

Genome source: UCSC Genome Browser (http://genome.ucsc.edu/). Maximum order considered for selection: 14; only one strand was used. Optimal order (AIC or BIC) generally increases with the length and the complexity of a genome.

suggests that the BPO order acts as a lower bound on the order to be used in such clustering exercises.

In the context of phylogeny reconstruction, Markov models are usually part of more complex methods that also capture the rate of substitution (10). Our results suggest that even simple Markov models, constructed using an appropriate order (i.e. BPO order in this case), can capture the structure of the phylogeny to a remarkable degree.

An application to metagenomics

Motivated by the success of our classifier for phylogeny reconstruction, we looked at the 'binning' problem faced in metagenomic contexts. Metagenomic studies typically explore the uncultured microbial world by shotgun sequencing DNA samples from various natural environments. The binning problem arises in the post-processing step, where the goal is to classify the sequence fragments taxonomically. Phylopythia (20) and phymm (21) are two popular methods developed for this purpose, which do not rely on sequence alignment. Phylopythia uses support vector machines (SVMs) based on frequencies of oligonucleotides of different sizes, whereas phymm uses interpolated Markov models to characterize the variable-length oligonucleotide frequencies specific to different taxa. The former has been shown to work well for reads of at least 1000 bp length, whereas the latter for lengths as low as 100 bp. We explored the possibility of whether the variable length oligonucleotide frequencies are really necessary, or using plain Markov models of appropriate but fixed order will work equally well.

We therefore used the script supplied by phymm that downloads all current bacterial and archaeal genomic and taxonomic data from RefSeq (22). This resulted in 1470 different genomes across 2 domains, 14 phyla, 21 classes, 39 orders and 27 genera, which contained genomic information for at least 10 species in each clade. We built Markov models of orders ranging from 0 to 10 for each clade at each taxonomic rank.

To evaluate the power of our models in describing taxa of known organisms and in determining the taxum of an unknown organism, we performed a standard 5-fold cross-validation test: Markov models were built using full genomic sequences of 4/5th of the species belonging to each clade at each taxonomic rank, and tested on the left out 1/5th as follows. To emulate real-life situations, instead of using the learned Markov models to score the full genomic sequences of the left out 1/5th species, we scored 10 randomly chosen fragments of lengths 100 and 1000 bp from the genomes. The clade corresponding to the Markov model that scored the fragment the highest was assigned to the fragment. The class-normalized accuracy was computed as described before (20) for each taxonomic rank and at each Markov order (Figure 2). Cladewise average sensitivity results are available in the Supplementary Results. The APO order for the full set was 12, whereas the BPO order was 9. The performance of the classifier was best at orders close to 9, indicating that the BPO order is indeed most informative in this context. Supplementary Figure S2 through Supplementary Figure S5 display the behavior of the sensitivity–specificity and precision–recall pairs attained at all orders, which further support this observation. Furthermore, this accuracy is comparable with that achieved by phylopythia and phymm (Table 2).

Motif discovery

The problem of motif discovery is encountered frequently in genomics and proteomics. Examples include finding the sequence specificity of a transcription factor (TF) from co-regulated regions, conserved patterns in a family of protein sequences, signals at splice junctions, etc. Computationally, the problem can be posed as identifying short overrepresented patterns within a set of biological sequences. A plethora of tools have been developed for this purpose over the years (23). Most of these methods assume that the input sequences can be described with a Markov model. Motifs that are most different from this 'background' Markov model are then identified using deterministic (24) or stochastic (25) approaches. While higher order Markov models have been shown to perform better in practice (17), most tools arbitrarily choose an order, typically 3 or 5, based on its performance on a few test cases. We systematically applied AIC- and BIC-based order selection methods to the background Markov model for yeast and human data sets generated through high-throughput chromatin immunoprecipitation (ChIP) experiments before performing motif discovery.

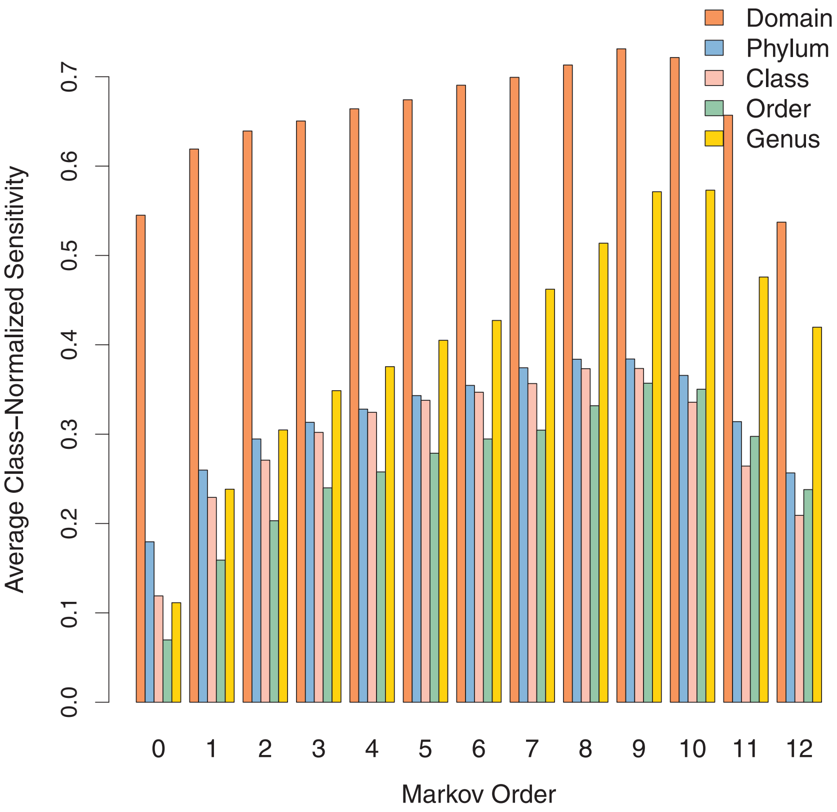

Figure 2. Class-normalized sensitivity at each taxonomic level for 100 bp fragments tested across classifiers built from 13 different Markov orders. Thirteen models were built during each of the 5 folds of cross-validation. The held out set of organisms was tested with each model ('Materials and Methods' section) to identify their taxa, by taking 10 random fragments of length 100 bp from their genome. The best average sensitivity for all five taxonomic levels is achieved at orders between 8 and 10. The BIC-predicted optimal order is equal to 9.

Table 2. Class-normalized accuracy of three taxa predictors in percentages

Rank	Markov (100 bp)	Markov (1000 bp)	Phylopythia (1000 bp)	Phymm (100 bp)
Domain	73.0 (2)	85.2 (2)	57.7 (3)	N/A
Phylum	38.1 (14)	56.8 (14)	40.6 (14)	36.7 (14)
Class	35.6 (21)	60.9 (21)	30.7 (22)	37.4 (21)
Order	31.3 (39)	60.2 (39)	6.4 (29)	32.8 (34)
Genus	48.1 (27)	75.0 (27)	4.4 (31)	25.0 (53)

The numbers in the parenthesis indicate the number of clades considered by the program. The highest accuracy (class-normalized sensitivity; Supplementary Methods) achieved for lengths 100 and 1000 bp using Markov models is shown in the first two columns. Class-normalized sensitivity as published by phylopythia for lengths 1000 bp and those computed from phymm (21**;** Supplementary Tables S7–S11**,** therein) for 100 bp are shown in the adjacent columns.

We used priority (26), a Gibbs sampling-based motif discovery program that allows users to define the background Markov model of their choice. While priority was originally developed to incorporate additional information in the form of positional priors, it can also be used with a non-informative uniform prior, which is how we employed it here. Use of EM-based tool meme (27) for low background orders led to results displaying a similar trend as priority (data not shown). However, meme does not appear to support large orders, since we encountered run-time errors for background model orders greater than 7.

A motif reported by priority was considered 'correct' if the Euclidean distance between the learned motif and the literature consensus motif was less than a predetermined cutoff (26,28). Being a stochastic algorithm, priority can yield different results during each run. To account for this variability, we ran priority on each data set 20 times and, following Gordân *et al.* (29), report the median number of successes for each order.

Motif discovery in yeast

We examined the ChIP-on-chip data published by Harbison *et al.* (28), which consists of several TFs profiled in multiple environment conditions. The APO and BPO orders for the sequence spotted on the microarray were 7 and 5, respectively.

To assess whether the APO or the BPO orders were informative background models for motif discovery, we built Markov models of order 0 through 9 for the same set. Using each of these models as the background model, we used priority to identify the most enriched motif in 156 sequence sets with known TF-specific binding motifs, i.e. motifs that have been characterized in the literature and used previously for assessing motif discovery methods (26,30). The performance of priority with background Markov orders from 0 through 9 is shown in Figure 3A.

Two different distance cutoffs, namely, 0.18 (as used by Harbison *et al.*) and 0.24 [as used by Linhart *et al.* (31)] yield similar trends: the maximum number of successes is achieved at order 5 or 6. This matches closely with the BPO order.

It is interesting to note that with 0.24 as the distance cutoff, we find well over 60 motifs correctly at the BPO background model order of 5. This is close to the best performance of priority with the inclusion of positional priors; i.e. about 70 correctly found motifs with 0.25 as the distance cutoff (29). This suggests that a combination of optimal background order with appropriate prior information may further enhance the performance of motif discovery methods.

Motif discovery in human promoters

To assess the role of order selection in motif discovery for more complex genomes, we examined the human promoter data set compiled by Linhart *et al.* (31). This data consist of 20 human promoter sets bound by TFs in different ChIP experiments conducted by multiple laboratories. The respective TF binding motifs are listed in transfac (32). To build background Markov models, we used the full human promoter set compiled by Linhart *et al.* The APO and BPO orders for this set turned out to be 9 and 7, respectively.

Figure 3B shows the number of motifs predicted correctly by priority in these 20 sequence sets. The order at which priority finds the maximum correct motifs at either cutoff is 7, which is the BPO order for the background model. Interestingly, orders 3, 4 and 5 that are the typical defaults in motif discovery tools give the worst results.

Functional classification of promoters

We next explored the possibility of building a Markov model-based classifier to predict the functionality of human promoters. We considered genes in the human genome that are *tissue specific*, i.e. specifically expressed in only one tissue. From gene expression data published by Su *et al.* (33), Schug *et al.* (34) report genes that are specifically expressed in one or few of the 25 tissues. Of these 25 tissues, 12 tissues had at least 50 genes expressing in only that tissue based on their recommended cutoff. These tissues included cerebellum, corpus callosum, cortex, heart, liver, lung, pituitary gland, placenta, spleen, testis, thymus and thyroid. We examined the 12 promoter sets corresponding to the 50 genes in each of these 12 tissues, and built Markov models for each promoter set. The BPO order for all tissues except liver was 3. In case of liver, the BPO order was 2, with 3 being a close second.

To evaluate the descriptive power of the Markov models, we built a multiclass classifier to distinguish promoters across these 12 sets. Each promoter sequence was scored using each of the 12 Markov models. The downstream gene was considered to be specific for the tissue for which the log-likelihood had the largest value.

We first applied this procedure to the training data itself; i.e. all the promoters on which the models were trained (Figure 4; gray curve, right scale). As expected, the accuracy of the classifier went up with the order of the Markov model, with all 600 promoters predicted correctly beyond the sixth order. This implies that more complex models fit the training data better.

To assess the performance of the classifier against unseen data, we used the standard 5-fold cross-validation

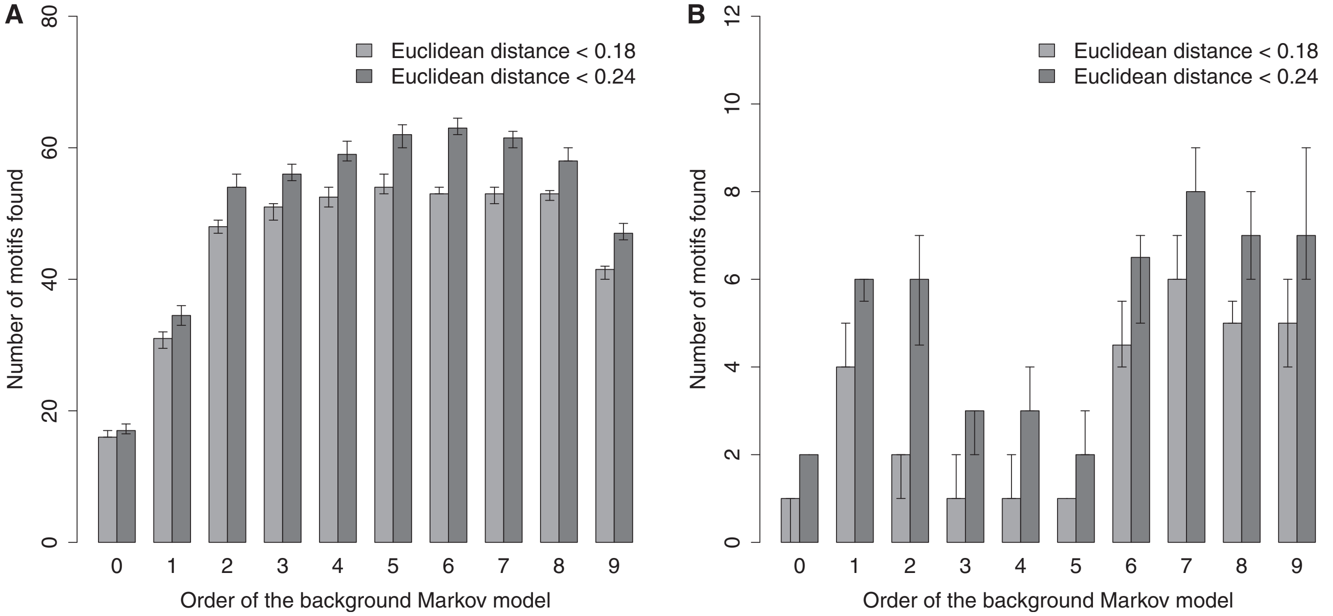

Figure 3. Number of motifs identified correctly by priority for yeast and human promoter data sets. Priority was run on each promoter set 20 times. A barplot of the number of times the returned motifs matched the literature consensus motif is shown here for each order. A match is determined by the condition that the Euclidean distance between a found motif and the literature consensus motif be less than a predetermined threshold; we used 0.24 and 0.18 as the thresholds. The highest number of matches occurs at (**A**) order 5 or 6 for the 156 yeast promoter sets (BIC-predicted optimal order = 5) and (**B**) order 7 for the 19 human promoter sets (BIC-predicted optimal order = 7).

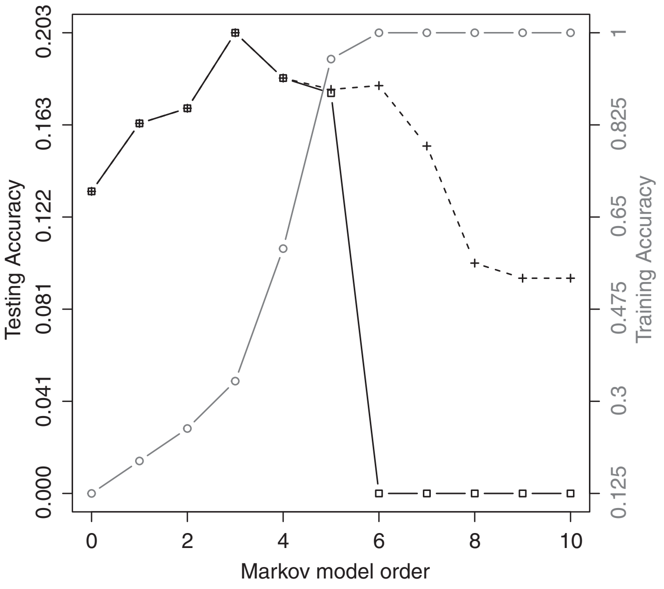

Figure 4. Accuracy of tissue-specificity classifiers based on Markov models of different orders. Here, accuracy is defined as the number of tissue-specific promoters predicted correctly by a classifier at a Markov order divided by the total number of promoters. Total number of tissues: 12, number of tissue-specific promoter sequences for each tissue: 50. On the full training set, the accuracy of the classifier increases with the order and reaches the maximum at order 6 (gray curve, right scale). On test data, however, the order 3 classifier performs the best (solid black curve, left scale), with the predictive power vanishing order 6 onward. Addition of a pseudocount while computing the probability distribution, as described in 'Materials and Methods' section, improves the performance of the classifier at higher orders (dashed curve, left scale), but cannot surpass the performance of the without pseudocount classifier at the BIC-predicted optimal order equal to 3.

procedure. That is, the model is trained on 4/5th of the sequences and evaluated on the left out 1/5th of the sequences; this whole procedure is carried out five times with each sequence being tested once. The results are shown in Figure 4 (solid black curve, left scale). The highest number of correctly classified promoters (122, corresponding to an accuracy of ~0.2) occurs for the BPO order, i.e. 3. Note that a non-informative classifier that assigns tissues to each promoter randomly will correctly predict an average of only 50 promoters. Indeed, the binomial *P*-value of predicting at least 122 promoters correctly for this data set, under the assumption of uniform probability of success across the 12 sets, is less than 10^{-20}.

The accuracy of our classifier drops sharply beyond order 5 primarily due to singularities in the probability matrix (Figure 4; solid black curve, left scale). This is an outcome of overfitting: many words of length ≥ 7 are not represented in the training set, but appear in the test sequences. We also, therefore, repeated our classification exercise after adding a pseudocount of 1 to the word counts, which is a standard practice for estimating probabilities when there are relatively few observations. With the pseudocount added (Figure 4; dashed curve, left scale), the accuracy of our classifier goes up significantly for orders greater than 4, but cannot match that of the BPO order. Interestingly, for orders ≥ 9, the accuracy of our classifier with pseudocount approaches that of a random-guess classifier; this point is further elaborated upon in 'Discussion' section.

DISCUSSION

In this study, we demonstrated that the performance of genomic sequence analysis methods employing Markov models is highly dependent on the order of the model used. We described the utility of two well-known model selection methods (AIC- and BIC based) for identifying the optimal order for Markov models in biological contexts. Although Markov models have been widely employed in modeling biological sequences, the issue of which order to use is typically ignored.

For instance, order 3 or 5 is used as defaults by many motif discovery tools. While these orders might work for less complex genomes, such as yeast, our results revealed that order 7 worked the best for human sequence data. Moreover, for both genomes (yeast and human), the BPO order led to the most accurate results. Interestingly, we found the first- and the second-order Markov models also to perform well for human TF motif discovery (Figure 3B). We believe that chromatin structure around a promoter might be involved here, as nucleosome positioning has been shown to have first-order correlations (35). This is an important implication of this study and needs to be investigated further.

We demonstrated that mammalian phylogeny can be reconstructed successfully using whole-genome Markov models of the BPO order. Admittedly, our algorithm for phylogenetic tree construction is primitive and does not have a clear interpretation in terms of evolutionary distances. However, the fact that it can learn true relationships across species has significant implications in the areas of phylogenetics and metagenomics. First, phylogenetic trees are typically built using multiple alignments (36) and more complex probabilistic approaches (10) that capture evolutionary rates across various parts of the genome. While such methods undoubtedly yield more informative trees, our results show that simple Markov models of appropriate orders can capture the structure of the phylogeny to a remarkable degree.

We also demonstrated that BPO models are useful for classification purposes in the context of metagenomics. Even for short fragments of length 100, a simple classifier built using these models lead to accuracies comparable to, if not better than, the more complex programs phylopythia and phymm. We note, however, that the actual data sets on which the three methods were tested are different: phymm is trained on more data than phylopythia, while our method is evaluated on more data than phymm. This changes the number of clades at each level and also the number of genomes in each clade.

Markov models have been used before as classifiers for identifying functional genomic regions; e.g. CpG islands and coding regions (10). However, to the best of our knowledge, the present study is the first attempt to implement a multiclass classifier to identify the functionality of a mammalian tissue-specific promoter. Specifically, we showed that the performance of simple Markov models of the BPO order is admirable in identifying which tissue a given promoter is most likely to be expressed in. Furthermore, Markov models are generative in nature; i.e. a new/synthetic sequence can be generated using the probability distribution defined by a Markov model. Our results therefore have significant implications in designing synthetic tissue-specific promoters.

The behavior of our classifier for tissue prediction deserves a deeper look for the case when pseudocount of 1 was added (Figure 4; dashed curve, left scale). Such classifiers predict around 1/12th of the promoters correctly for large Markov orders. A random-guess procedure will result in a similar accuracy; e.g. by rolling a dodecahedron-shaped fair die, and assigning a tissue label corresponding to one of the 12 numbers that shows up. In terms of Bayesian statistics, addition of the same pseudocount for all words is equivalent to incorporating a flat prior during parameter estimation (37). As the data, i.e. word counts, become sparse at high Markov orders, the uniform prior starts dominating, making the classifier behave similar to a random-guess classifier. Middle-order Markov models (orders 5–7) benefit the most from the addition of a pseudocount, because it helps to regularize the probability distributions; however, their accuracy never reaches that of the BPO order. Since the BIC already incorporates a strong penalty for model complexity, it indirectly penalizes models based on sparse word counts and obviates the incorporation of pseudocounts.

We noted earlier that the accuracy of our classifier for the metagenomic binning problem (Table 2) is comparable with that achieved by phylopythia (20) and phymm (21). This behavior can be understood as follows. Phylopythia builds SVMs using frequencies of oligonucleotide of lengths between 2 and 6. Going to larger lengths result in an exponential increase in the size of the parameter space, making SVM-based learning infeasible. It is likely that our simpler method outperforms phylopythia because it can capture information in longer oligonucleotides. Phymm, in contrast, also has a more complex model, but includes models of lengths up to 12. We suspect that the oligonucleotides of these larger lengths are primary contributors to the success of phymm.

Table 1 shows that for the same sequence data, the APO order is generally larger than the BPO order. In other words, the BIC order selection procedure tends to select simpler models with greater parsimony. This is expected, as indicated by Equations (1) and (2) ('Materials and Methods' section): the BIC procedure puts a heavier penalty on model complexity. Similar differences in the APO and BPO orders for the same sequence data are also seen in other genomic sequence analysis contexts explored in this study.

Before we conclude, we make two methodological remarks.

First, an alternate model selection criterion called AIC$_c$ (38,39) is often recommended in place of AIC. In fact, AIC$_c$ is AIC corrected for data sizes that are small relative to the number of parameters in the model. In genomic contexts, the small data size problem is expected to show up at sufficiently high Markov orders. However, for the data we used in this article, order selection based on the AIC$_c$ did not lead to optimal orders that were much different from the APO orders (Supplementary Table S1). Specifically, for the motif discovery and phylogeny background data, the APO and AIC$_c$ optimal orders turned out to be identical. For the metagenomics data, the AIC$_c$ optimal orders were not too different from the APO orders. In comparison, the BPO orders were, by and large, 2 or 3 less than the APO orders. This implies, *post facto*, that the choice between AIC and AIC$_c$ is perhaps not relevant in the genomic contexts we have explored. From a fundamental viewpoint, the form of the small sample correction to AIC is not universal because it inherently depends on the class of models and the nature of noise in the data. Therefore, it is not clear if the standard AIC$_c$ form, which was originally derived for linear models with normal noise, continues to be either valid or useful for Markov models. While (38) recommend using a small sample corrected form of AIC$_c$ instead of AIC, other experts (39) recommend using this particular AIC$_c$ form with caution outside of the realm of models for which it has been explicitly derived or demonstrated.

Second, it is important to note that while AIC and BIC scores [Equations (1) and (2) in 'Materials and Methods' section) appear to have similar mathematical forms, they originate from entirely different considerations. The AIC procedure, which does not assume that the true but unknown model is necessarily included in the set of models considered, attempts to find the model within this set that is closest to the true model that generated the data. The AIC procedure therefore tries to minimize the prediction error over the set of models, but is also known to be *inconsistent*; i.e. the APO order may overshoot the true order with non-zero probability (40). The BIC procedure, on the other hand, is designed to find the true model under the assumption that the true model is included in the set of models considered, and is known to be strongly consistent (40). Therefore, the model selection method itself needs to be chosen with reference to the context and goals of the downstream analysis (41).

In all the applications explored here, however, the BIC clearly produces results that are optimal from a biological perspective. Given a set of input sequences, computing the BPO order is relatively straightforward. We conclude that this small additional step before embarking upon motif discovery, phylogeny reconstruction, taxum identification in metagenomics or eukaryotic tissue-specific promoter modeling, can lead to remarkable improvement in the outcome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–5, Supplementary Methods, Supplementary Results and Supplementary Data sets 1–11.

ACKNOWLEDGEMENTS

The authors thank Sameet Mehta for many discussions during initial stages of this work and Ramon van Handel for clarifying certain statistical subtleties regarding AIC and BIC. The authors also thank two anonymous referees for insightful suggestions.

FUNDING

Department of Biotechnology, Government of India [BT/PR7594/BRB/10/490/2006 to S.G. and M.A.]; Department of Science and Technology, Government of India [Ramanujan Fellowship to L.N.]. Funding for open access charge: Indian Institute of Science Education and Research, Pune, India.

Conflict of interest statement. None declared.

REFERENCES

1. Dehnert,M., Plaumann,R., Helm,W.E. and Hutt,M.T. (2005) Genome phylogeny based on short-range correlations in DNA sequences. *J. Comput. Biol.*, **12**, 545–553.
2. Yoseph,B., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein-DNA binding sites. In: *RECOMB'03*. ACM Press, NY, USA, pp. 28–37.
3. Istrail,S. and Davidson,E.H. (2005) Logic functions of the genomic cis-regulatory code. *Proc. Natl Acad. Sci. USA*, **102**, 4954–4959.
4. Vaillant,C., Audit,B. and Arneodo,A. (2007) Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.*, **99**, 218103.
5. Kumar,P.P., Mehta,S., Purbey,P.K., Notani,D., Jayani,R.S., Purohit,H.J., Raje,D.V., Ravi,D.S., Bhonde,R.R., Mitra,D. *et al.* (2007) SATB1-binding sequences and alu-like motifs define a unique chromatin context in the vicinity of human immunodeficiency virus type 1 integration sites. *J. Virol.*, **81**, 5617–5627.
6. Zhang,M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.*, **7**, 919–932.
7. Alexander,R.P., Fang,G., Rozowsky,J., Snyder,M. and Gerstein,M.B. (2010) Annotating non-coding regions of the genome. *Nat. Rev. Genet.*, **11**, 559–571.
8. Walsh,J.B. (2001) Genome evolution: overview. In: *Encyclopedia of Life Sciences*. John Wiley and Sons Ltd, NJ, USA.
9. D'haeseleer,P. (2006) How does DNA sequence motif discovery work? *Nat. Biotechnol.*, **24**, 959–961.
10. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.
11. Lomsadze,A., Ter-Hovhannisyan,V., Chernoff,Y.O. and Borodovsky,M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.*, **33**, 6494–6506.
12. Hansen,N.R. (2009) Statistical models for local occurrences of RNA structures. *J. Comput. Biol.*, **16**, 845–858.
13. Wu,T.J., Hsieh,Y.C. and Li,L.A. (2001) Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. *Biometrics*, **57**, 441–448.
14. Dai,Q., Yang,Y. and Wang,T. (2008) Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics*, **24**, 2296–2302.
15. Chang,G. and Wang,T. (2011) Weighted relative entropy for alignment-free sequence comparison based on Markov model. *J. Biomol. Struct. Dyn.*, **28**, 545–555.
16. Thakur,V., Azad,R.K. and Ramaswamy,R. (2007) Markov models of genome segmentation. *Phys. Rev. E*, **75**, 011915.
17. Thijs,G., Lescot,M., Marchal,K., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2001) A higher-order background model improves the detection of potential promoter regulatory elements by Gibbs sampling. *Bioinformatics*, **17**, 1113–1122.
18. Akaike,H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **19**, 716–723.
19. Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.
20. McHardy,A.C., Martin,H.G., Tsirigos,A., Hugenholtz,P. and Rigoutsos,I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*, **4**, 63–72.
21. Brady,A. and Salzberg,S.L. (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods*, **6**, 673–676.
22. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. *et al.* (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16.
23. Das,M.K. and Dai,H.K. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8(Suppl 7)**, S21.
24. Dempster,A., Laird,N. and Rubin,D. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B*, **39**, 1–38.
25. Gelfand,A. and Smith,A. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, **85**, 398–409.
26. Narlikar,L., Gordân,R. and Hartemink,A.J. (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, **3**, e215.
27. Bailey,T. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Intelligent Systems for Molecular Biology*. AAAI Press, CA, USA, pp. 28–36.
28. Harbison,C., Gordon,D., Lee,T., Rinaldi,N., Macisaac,K., Danford,T., Hannett,N., Tagne,J., Reynolds,D., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
29. Gordân,R., Narlikar,L. and Hartemink,A.J. (2010) Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res.*, **38**, e90.
30. Georgiev,S., Boyle,A., Jayasurya,K., Ding,X., Mukherjee,S. and Ohler,U. (2010) Evidence-ranked motif identification. *Genome Biol.*, **11**, R19.
31. Linhart,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.
32. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
33. Su,A.I., Cooke,M.P., Ching,K.A., Hakak,Y., Walker,J.R., Wiltshire,T., Orth,A.P., Vega,R.G., Sapinoso,L.M., Moqrich,A. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
34. Schug,J., Schuller,W., Kappen,C., Salbaum,J., Bucan,M. and Stoeckert,C. (2005) Promoter features related to tissue specificity as measured by shannon entropy. *Genome Biol.*, **6**, R33.
35. Segal,E., Fondufe-Mittendorf,Y., Chen,L., Thåström,A., Field,Y., Moore,I., Wang,J. and Widom,J. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
36. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
37. Strelioff,C.C., Crutchfield,J.P. and Hübler,A.W. (2007) Inferring markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Phys. Rev. E (Statistical, Nonlinear, and Soft Matter Physics)*, **76**, 011106.
38. Burnham,K.P. and Anderson,D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer, New York.
39. Claeskens,G. and Hjort,N.L. (2008) *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.
40. Katz,R.W. (1981) On some criteria for estimating the order of a Markov chain. *Technometrics*, **23**, 243–249.
41. Zucchini,W., Claeskens,G. and Nguefack-Tsague,G. (2010) Model selection. In: Nair,V. *et al.* (eds), *International Encyclopedia of Statistical Sciences*. Springer, Berlin, Germany.