Mathematical Modelling and Analysis
Volume 13 Number 1, 2008, pages 135–144
© 2008 Technika ISSN 1392-6292 print, ISSN 1648-3510 online

LOGIT ANALYSIS OF GENETIC DATA

J. ŽIDANAVIČIŪTĖ

Vilnius Gediminas Technical University
Saulėtekio al. 11, LT-10223 Vilnius, Lithuania
E-mail: jurz@fm.vgtu.lt

Received November 11, 2007; revised December 20, 2007; published online February 15, 2008

Abstract. A new framework of genetic sequence statistical analysis based on generalized logit model is introduced. Logit analysis is applied to assess the dependence structure (interactions) between DNA nucleotides and to test hypothesis about Markov order of these dependencies. The procedure proposed seeks the non-coding subsequences which are homogeneous but yet non-Markov. It has been shown, that even homogeneous DNA regions can not be treated as the first order Markov sequences.

Key words: DNA, generalized logit, Markov chain, log-linear model, re-sampling.

1 Introduction

Categorical data arise from different sampling frameworks. The goal of statistical analysis is to find a dependence structure between a set of categorical variables. There are various models available for describing the nature of the association between these variables. In this study the logit analysis is applied to assess the dependence structure (interactions) between DNA nucleotides and to test hypothesis about Markov order of these dependencies.

The problem is closely related to the context-dependent evolutionary model of DNA sequences. It is known [13] that for the time-reversible evolution, the stationary distribution of nucleotides in DNA sequences inherits its finite order Markov dependence structure from local interactions of the nearest nucleotides in evolutionary process. If only the first nearest neighbours of a nucleotide are involved in its mutation process (this is a quite natural assumption for non-coding DNA sequences) the stationary distribution of nucleotides of time-reversible evolution is the first order Markov chain. Thus, the higher order Markov dependencies in non-coding DNA sequences indicates that the evolution is probably not time-reversible.

Many investigations of DNA sequences are devoted to so-called long-range dependence phenomena (see, for instance [18]). The term came from the theory of stationary time series. It is known that it can be a consequence of different patterns of (probabilistic) inhomogeneity of a observed sequences. Thus, another issue addressed in the paper is the question if the higher order of Markov property is related to inhomogenity of DNA sequences.

The basics of DNA sequences and a brief survey of relevant literature are presented in Section 2. In Section 3 a representation of Markov chain (field) by conditional odds is given and appropriate structure of genetic data is introduced. This ensures that the classical assumptions of the generalized logit model are fulfilled. The hypotheses of Markov property and equivalence of DNA strands are formulated as submodels of the basic model. Section 4 contains results of statistical analysis of bacterial genome taken from the *Genbank* database.

2 Markov Chain Modeling of Genetic Sequences

2.1 Basics of DNA sequences

DNA sequences are long sequences of nucleotides (nitrogenous bases). At each position it has one of the nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). Thus, a DNA (or nucleotide) sequence can be viewed as a sequence of categorical random variables taking their values from a finite alphabet $\mathcal{A}$ with four letters $\{A, T, C, G\}$.

It is convenient to transform a DNA sequence to a two dimensional binary sequence according to two physical properties of nucleotides: purine-pyrimidine and the number of bounds, either two or three.

A gene is a protein coding nucleotide sequence, and DNA sequences located between genes are called non-coding genome sequences. In our study we use non-coding regions of DNA sequences as being more likely to be homogeneous.

Typical models for DNA sequences are homogeneous m-order Markov chains (models Mm) on a finite alphabet (state space) $\mathcal{A}$. These models represent the local dependencies in the sequence: the probability of occurrence of a letter at a given position depends only on the m previous letters in the sequence (and not on the position).

The first stochastic models of DNA sequence evolution in time assume that it is a homogeneous Markov process and nucleotides in the DNA sequence evolve independently from each another. Recently context-dependent evolution models have been proposed (see [13] and references therein). In these models mutations of each nucleotide depend on m its nearest-neighbouring nucleotides from the each side (context). Usually $m = 1, 2, 3$.

If evolution of DNA sequence in time is reversible and mutation of nucleotides depends on their nearest neighbours, say m from the each side, then the stationary distribution of nucleotides in the sequence is mth order Markov chain. Thus, a topical problem is to estimate the order of Markov chain model in DNA sequences.

2.2 Estimating of Markov model

Estimating of Markov model means to choose the Markov order m and to estimate transition probabilities

$$\Pi := \bigl\{\pi(a, a_{m+1}),\ a = (a_1, \dots, a_m) \in \mathcal{A}^m,\ a_{m+1} \in \mathcal{A}\bigr\}.$$

The conventional model for a mth order Markov chain has $(n-1)n^m$ model parameters with $n = |\mathcal{A}|$ being the number of elements in $\mathcal{A}$. The major problem in statistical analysis of such models is that the number of parameters (the transition probabilities) increases exponentially fast with respect to the order m of the model. This large number of parameters discourages researches from using a higher-order Markov chain directly.

A. Raftery [16] have proposed a higher-order Markov chain model which involves only one additional parameter for each extra lag after the first one and proved that the autocorrelations satisfy a system of linear equations similar to the Yule-Walker equations. His model have been extended and refined in [7] and [6].

A more sophisticated approach to decreasing of the number of model parameters has been presented in [5]. The idea of *variable length Markov chain* (VLMC) model is that the memory length (depth) of the sequence at any point in time is allowed to depend on the preceding history which is represented via *context tree*. In this case order selection problem is stated as estimating problem of the context tree (or the interaction structure) of the sequence.

Eggar [12] have proposed a procedure of testing whether, given a sequence data, there is a Markov chain of the first order that is a likely model to fit it. The test is suggested by geometric arguments (graph theory).

The problem of estimating the Markov order or, more generally, the context tree has been addressed in many papers, see [8, 9, 14, 15, 17, 19] and references therein. In [8] and [9] the strong consistency of the BIC estimator, without any prior bound on memory length,of the Markov order and the context tree, respectively, is demonstrated.

The problem is really difficult: the Bayesian estimator or minimum description length (MDL) estimator, of which the BIC estimator is an approximation, is shown [8] to be inconsistent for the uniformly distributed i.i.d. sequence. DNA sequence order estimators for fixed and variable length Markov models and its practical performance have been studied in [11]. However, the gain in using VLMC models compared to fixed order Markov chains for DNA sequences, at least for their classification, is small [10].

Usually, it is supposed (and this applies to all earlier cited references) that the Markov chain is homogeneous. It seems that for DNA sequences this assumption does not hold in general.

In a sense, this paper is inspired by investigations of Avery and Henderson [3] (see also [2]). Supposing that DNA sequences are generated by a homogeneous Markov chain of the order m, they have used log-linear model to estimate m. Their study shows that $m = 1$ does not provide a good fit to the data while the null hypothesis of the second-order Markov chain is not rejected.

3 Markov Property and Generalized Logit

We start with basic notions of discrete finite-state Markov fields (chains). Then a special structure of genetic data is introduced which is convenient for logit analysis and testing of the probabilistic equivalence of the two complementary strands in DNA helix.

3.1 Markov property and conditional odds

Let $\mathcal{A}$ be a finite alphabet with $|\mathcal{A}| = \mathrm{card}\mathcal{A}$ elements and $N := \{1, \ldots, n\}$. Fix some positive integer $m < n/2$ and define the "interior" N° and "boundary" ∂N of N,

$$N^\circ := (m+1, \ldots, n-m), \quad \partial N := N \setminus N^\circ,$$

and a collection of neighbourhoods

$$U(\ell) = U_m(\ell) := [\ell - m, \ell + m] \setminus \{\ell\}, \quad \ell \in N^\circ.$$

Here $[i,j] := (i, i+1, \ldots, j),\ i < j,\ i,j \in N$, is an interval of integers. Given $x \in \mathcal{A}^n$ and a set of indices $I \subset N$, let $x_I := (x_i, i \in I)$ denote the corresponding subsequence of x.

Definition 1. A random sequence $x \in \mathcal{A}^n$ is a homogeneous Markov random field (chain) of order m (denoted by Mm) if $\forall \ell \in N^\circ$ and $a \in \mathcal{A}^n$

$$\mathbf{P}\{x_\ell = a_\ell | x_j = a_j, j \neq l\} = \mathbf{P}\{x_\ell = a_\ell | x_{U_m(\ell)} = a_{U_m(\ell)}\} =: p(a_\ell | a_{U_m(\ell)}). \tag{3.1}$$

Definition 2. Conditional odds $O_{y|b}(z)$ of y versus b given the values $z = x_{U_m(\ell)}$ of the m nearest neighbours $U_m(\ell)$ of $\ell \in N^\circ$, for some reference value $b \in \mathcal{A}$ with $p(b|z) > 0$, is the ratio

$$O_{y|b} = Q_{y|b}(z) := \frac{p(y|z)}{p(b|z)}, \quad y \in \mathcal{A},\ z \in \mathcal{A}^{2m}, \tag{3.2}$$

where the probabilities $p(y|z)$ are introduced in (3.1). Suppose that values of the Markov chain x are fixed on the boundary ∂N, $x_{\partial N} = c_{\partial N}$ for some $c \in \mathcal{A}^n$, and set

$$\mathcal{X}_+ := \{a \in \mathcal{A}^n :\ a_{\partial N} = c_{\partial N}\}.$$

Hamersley-Clifford theorem [4] implies the following statement.

Proposition 1. *If* $\mathbf{P}\{x = a\} > 0$ *for all* $a \in \mathcal{X}_+$, *the (conditional) distribution of the homogeneous Markov random field (chain)* Mm *is uniquely determined by the conditional odds* $O_{y|b}(z),\ y \in \mathcal{A},\ z \in \mathcal{A}^{2m}$, *for some reference value* $b \in \mathcal{A}$, *and there exists a function* $\lambda_m \colon \mathcal{A}^{m+1} \to \mathbf{R}$ *such that for each* $a = (a_1, \ldots, a_{2m+1}) \in \mathcal{A}^{2m+1}$

$$\log\Big(Q_{a_{m+1}|b}(a_{U_m(m+1)})\Big) = \sum_{j=1}^{m+1} \Big[\lambda_m\Big(a_{[j,m+j]}\Big) - \lambda_m\Big(a^{(b)}_{[j,m+j]}\Big)\Big], \tag{3.3}$$

where $a^{(b)} = \big(a_1, \ldots, a_m, b, a_{m+2}, \ldots, a_{2m+1}\big)$.

This means that a way to identify the model Mm is to determine its conditional odds (3.2).

Let us introduce the following structure of the observed sequence $x \in \mathcal{A}^n$ of the length $n = n_m(m+1) + m$, the quantity n_m being an integer. Set

$$X := \{(y_\ell, z_\ell), l \in S\}, \quad S = S_{n,m} = \{m+1, 2(m+1), \dots, n-m\},$$

where $y_\ell := x_{(m+1)\ell}$ is a target variable and $z_\ell = x_{U_m(\ell)}$ is a vector of explanatory variables, $\ell \in S$.

Let us assume that

(A1) $\{y_\ell,\ \ell \in S\}$ are conditionally independent given $\{z_j,\ j \in S\}$,

(A2) the conditional probabilities $\mathbf{P}\{y_\ell = a | z_j,\ j \in S\},\ a \in \mathcal{A}$, do not depend on the position $\ell \in S$.

Remark 1. Note that these assumptions are fulfilled if x is generated by a homogeneous Markov chain of the order m. The Markov property implies the additional conditions on odds (3.2). Namely, as (3.3) shows the odds depend on $y = x_\ell$ and $z = x_{U_m(\ell)}$ only through interactions x_{I_j}, $j = 0, \dots, m$, where $I_j = U_m(\ell) \cap [\ell - m + j, \ell + j]$. This gives a basis for testing Markovity and selection of the Markov order (see next subsection).

Remark 2. Assumptions (A1) and (A2) ensure that common conditions of the generalized logit model are satisfied [1]. The generalized logit is a regression-type model, i.e. a *conditional* model with given values of the explanatory variables z. This means that the probabilistic model of the explanatory variables $\{z_\ell,\ l \in S\}$ is not specified and can be treated as a nuisance nonparametric component of the model. This also means that the conditions of Markov property mentioned in Remark 1 are necessary but not sufficient.

3.2 Generalized logit

Generalized logit model is in fact a loglinear model for conditional odds (3.2)

$$\log\big(Q_{y|b}(v,w)\big) = \lambda\big(v,y,w\big), \quad y \in \mathcal{A},\ v,w \in \mathcal{A}^m. \tag{3.4}$$

The function $\lambda\colon \mathcal{A}^{2m+1} \to \mathbf{R}$ satisfies $\lambda(\cdot, b, \cdot) = 0$ and in general case (linearly) depends on $K = (|\mathcal{A}| - 1)|\mathcal{A}|^{2m}$ parameters. Model (3.4) is referred to as a *saturated* generalized logit model.

In view of (3.2), (3.4), and assumptions (A1), (A2), the conditional distribution of $\{y_\ell,\ \ell \in S\}$, given values of $\{z_\ell,\ \ell \in S\}$, is represented as a product of multinomial distributions. Counts $N(a)$ of occurrence of "words" $a \in \mathcal{A}^{2m+1}$ in the sequence $\{x_{[\ell-m,\ell+m]},\ \ell \in S\}$,

$$N(a) := |\{j \in S\colon\ x_{[j-m,j+m]} = a\}|, \quad a \in \mathcal{A}^{2m+1},$$

constitute a sufficient statistics for the underlying model. In case of Mm model the function λ given in (3.3) takes the following form

$$\lambda(a) := \lambda\big(a_{[1,m]}, a_{m+1}, a_{[m+1,2m+1]}\big) = \lambda_{\mathrm{Mm}}(a), \tag{3.5}$$

$$\lambda_{\mathrm{Mm}}(a) := \sum_{j=1}^{m+1} \Big[\lambda_m\Big(a_{[j,m+j]}\Big) - \lambda_m\Big(a^{(b)}_{[j,m+j]}\Big)\Big], \quad a \in \mathcal{A}^{2m+1}, \tag{3.6}$$

and depends on $k = (|\mathcal{A}|-1)|\mathcal{A}|^m$ parameters. Let us consider the case $m = 1$. Then generalized logit (3.4) in general case (saturated model) and in case of Markov model Mm (3.5), (3.6) takes the following simple forms

$$\begin{aligned} \log\big(Q_{y|b}(v,w)\big) &= \lambda_0(y) + \lambda_L(v,y) + \lambda_R(y,w) + \lambda_{LR}(v,y,w), \quad (3.7)\\ \log\big(Q_{y|b}(v,w)\big) &= \lambda_0(y) + \lambda_1(v,y) + \lambda_1(y,w), \quad v,y,w \in \mathcal{A}, \end{aligned}$$

respectively, where all these functions λ vanish provided any of their arguments take the reference value b.

Thus, the null hypothesis for the first order homogeneous Markov chain M1 is given by

$$H_0: \ \lambda_{LR}(v,y,w) = 0, \ \lambda_L(v,w) = \lambda_R(v,w) \quad \forall\ v,y,w \in \mathcal{A}. \tag{3.8}$$

Generalized logit model (3.7) is convenient for characterization of the probabilistic equivalence of the two complementary strands in the DNA helix. Taking into account the two basic properties of nucleotides mentioned in subsection 2.1, it is natural to define the one-to-one mapping $h\colon \mathcal{A} \to \mathcal{A}_1 \times \mathcal{A}_1,\ \mathcal{A}_1 := \{0,1\}$, by equality

$$\{h(A), h(C), h(G), h(T)\} = \{(0,0), (1,1), (0,1), (1,0)\}.$$

Let $h(a) = (h(a_1), \dots, h(a_k)),\ a \in \mathcal{A}^k, k = 2,3,\dots$, and let

$$h^*(a) = (h^*(a_k), \dots, h^*(a_1))$$

denote a complementary mapping for $a \in \mathcal{A}^k,\ k = 1,2,\dots$, where $h^*(a) = (1-z_1, z_2)$ provided $h(a) = (z_1, z_2)$, $a \in \mathcal{A}$. Since the code of the complementary DNA strand is running in a opposite direction, the stated probabilistic equivalence of the DNA strands means that

$$\mathbf{P}\{h(x) = h(a)\} = \mathbf{P}\{h(x) = h^*(a)\} \quad \forall a \in \mathcal{A}^n.$$

In our framework this implies that

$$\lambda\big(v,y,w\big) = \lambda\big(w^*, y^*, v^*\big) \tag{3.9}$$

for all $y, y^* \in \mathcal{A}$ and $v, w, v^*, w^* \in \mathcal{A}^m$ satisfying $h(y) = h^*\big(y^*\big)$, $h(v) = h^*\big(v^*\big)$, and $h(w) = h^*\big(w^*\big)$. Equality (3.9) determines equations of the null hypothesis about the equivalence of the DNA strands in the setting of saturated generalized logit model (3.9). For mth order Markov models with λ given by (3.5) and (3.6), the null hypothesis of the equivalence can be expressed similarly in terms of the function λ_m.

4 Statistical Analysis

We apply the logit analysis to assess the Markov property of genetic sequences of bacterial genome. DNA sequences of three bacteria, *Esherichia coli (ID=5)*, *Bordetella bronchiseptica(ID=1)* and *Coxiella burnetti(ID=4)*, from the GenBank database are used in the analysis. SAS software (procedures CATMOD, LOGISTIC, NLMIXED) is used for statistical analysis of the data.

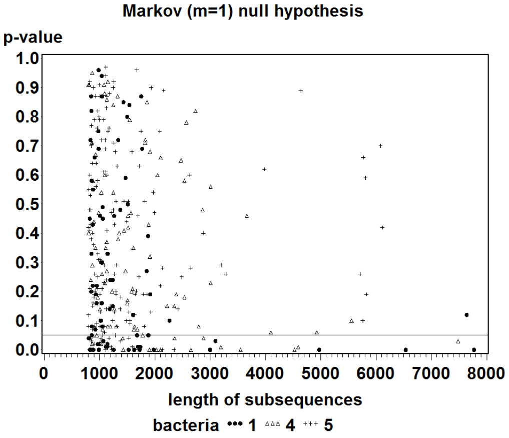

Figure 1. p-value for hypothesis about first order Markov chain for each subsequence.

4.1 Inhomogenity of data

The DNA sequences are known to be rather inhomogeneous. Therefore only non-coding regions of bacterium genome are considered since for primitive organisms they seem to have no direct impact on their vitality and survival. Moreover, the requirement of homogeneity of Markov chain is relaxed to some extent by replacing the loglinear model used by Avery and Henderson (Avery and Henderson (1999)) with more flexible generalized logit model and by omitting the second equality in the Markov null hypothesis (3.8):

$$H_0: \ \lambda_{LR}(v,y,w) = 0 \quad \forall \ v,y,w \in \mathcal{A}. \tag{4.1}$$

Nevertheless the null hypothesis (4.1) is rejected for whole non-coding DNA sequence ($p < 0.0001$) and for some of non-coding subsequences of the each bacteria (Fig.1). For the whole non-coding sequences the null hypothesis of the Markov order $m = 2$ and $m = 3$ are rejected as well with p-value $p < 0.0001$. Notice that Markov model of the order m=3 has 624 parameters.

In order to investigate the level of inhomogenity of the non-coding DNA sequences in the setting determined by assumptions (A1), (A2) and generalized logit model (3.7), the logistic regression model with a binary response variable y, ($y = 1$, if nucleotide have 3 bonds and $y = 0$, if nucleotide have 2 bonds) is fitted separately for each non-coding sequence of the length greater than 800 of the each bacterium.

These results demonstrate that the even non-coding sequences of the same bacteria is rather inhomogeneous (Fig.2). A standard way to deal with inhomogenity in data is to apply mixed models. SAS procedure NLMIXED was applied to fit the logistic model for y with an identifier (the running number) of the non-coding subsequence as a subject for the random effect. Unfortunately, the procedure NLMIXED failed to fit the model, probably, because of large number of parameters, great extent and complexity of the problem, and inhomogeneity of the data which mismatches with the underlying assumptions of mixed models. Therefore a direct method described in the next subsection is applied to deal with the data inhomogeneity.

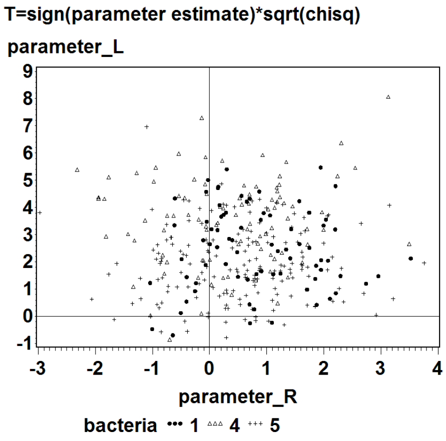

Figure 2. Example of inhomogeneity of the non-coding DNA sequences by using estimates of two parameters. In the case of homogeneity values of statistic T should be between -3 and 3.

4.2 Markov property of homogeneous DNA sequences

The goal of the study presented in this subsection is to check if the violation of the assumptions of the model M1 is caused merely by the data inhomogeneity.

The procedure proposed seeks the non-coding subsequences which are homogeneous but yet non-Markov. More precisely, it seeks subsequences for which inhomogeneity is less expressed than the non-Markov dependence structure.

The procedure is described in three steps: **I.** Several the longest non-coding sequences (of each bacterial), for which the model M1 is rejected, are used for analysis. They are divided into subsequences of approximately the same length ranging from 712 to 848. **II.** The subsequences for which the model M1 is rejected are selected for further analysis. **III.** The elements of the data set $\{(y_j, z_j), j \in S\}$ (the triplets) of selected subsequences are randomly divided into 10 disjoint groups of approximately equal sizes and the homogenity of the groups is tested within the generalized logit setting determined by (A1), (A2), and (3.7). The test of homogenity is based on resampling.

The subsequences for which the null hypothesis of homogenity of the random groups is not rejected are treated as homogeneous. These sequences are exactly what we are seeking for. The results demonstrate (see Table 1), that approximately 41.37% of subsequences are inhomogeneous.

Table 1. The results of homogenity of subsequences.

Bacterial	COL1	COL2	COL3
Esherichia coli	14	6 (42,8 %)	3 / 6 (50 %)
Coxiella burnetti	35	9 (25,7 %)	3 / 9 (33 %)
Bordetella bronchiseptica	37	14 (37,8 %)	6 / 14 (42,85 %)
Total	**86**	**29 (33,7 %)**	**12 /29 (41,37 %)**

COL1-number of subsequences
COL2-number of subsequences, for which H_0 about the model M1 was rejected
COL3-number of subsequences, for which homogenity hypothesis was rejected

5 Concluding Remarks

A new framework of genetic sequence statistical analysis based on assumptions (A1), (A2), and generalized logit model is introduced. It is directly related to finite-state Markov field specification and thus convenient for statistical analysis of Markov dependence structure. Furthermore, the generalized logit model is more flexible than alternatively used unconditional loglinear models and allows some extend of inhomogenity in the sequences under investigation. It is natural to suppose that the inhomogenity of DNA sequences causes the insufficiency for them of the first order Markov model M1, for example. A procedure based on resampling is performed to check if insufficiency of the Markov model M1 is merely a consequence caused by the data inhomogenity. The results obtained show that, within the generalized logit framework, DNA sequences are rather inhomogeneous and this can lead to their nonMarkovity. On the other hand, it is found that a significant part of nonMarkov sequences (approximately 58.63%) are homogeneous. Thus, even homogeneous DNA regions can not be treated as the first order Markov sequences.

References

[1] A. Agresti. *Categorical Data Analysis.* John Wiley & Sons, New York.

[2] P.J. Avery. Fitting interconnected Markov chain models – DNA sequences and test cricket matches. *The Statistician*, **51**(2):267–278, 2002.

[3] P.J. Avery and Daniel A. Henderson. Fitting Markov chain models to discrete state series such as DNA sequences. *Appl. Statist.*, **48**(1):53–61, 1999.

[4] J. Besag. Spacial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Ser. B*, **36**:192–236, 1974.

[5] P. Bühlmann and A.J. Wyner. Variable length Markov chains. *The Annals of Statistics*, **27**:480–513, 1999.

[6] W. Ching, Michael Ng and Shuqin Zhang. On computation with higher-order Markov chains. *Current Trends in High Performance Computing and Its Applications*, **1**:15–24, 2005.

[7] W.K. Ching, Eric S. Fung and Michael K. Ng. Higher-order Markov chain models for categorical data sequences. *Naval Research Logistics (NRL)*, **51**(4):557–574, 2004.

[8] I. Csiszár and Paul C. Shields. The consistency of the BIC Markov order estimator. *Ann. Statist.*, **28**(6):1601–1619, 2000.

[9] I. Csiszár and Zs. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory*, **52**:1007–1016, 2006.

[10] D. Dalevi, D. Dubhashi and M. Hermansson. Bayesian classifiers for detecting HGT using fixed and variable order Markov models of genomic signatures. *Bioinformatics*, **22**(5):517–522, 2006.

[11] D. Dalevi, D. Dubhashi and M. Hermansson. A new order estimator for fixed and variable length Markov models with applications to DNA sequence similarity. *Stat. Appl. Genet. Mol. Biol.*, **5**, 2006.

[12] M.H. Eggar. Validity of fitting a first order Markov chain model. *The Statistician*, **51**(2):259–265, 2002.

[13] J.L. Jensen. *Context dependent DNA evolutionary models*, volume 458. 2005.

[14] G. Morvai and B. Weiss. Order estimation of Markov chains. *IEEE Transactions on Information Theory*, **51**:1496–1497, 2005.

[15] Y. Peres and Shields P. Two new Markov order estimators. *arXiv:math.ST/0506080*, 2005.

[16] A. Raftery. A model for high-order Markov chains. *J. Roy. Statist. Soc.*, **B 47**:528–539, 1985.

[17] C.C. Strelioff, J.P. Crutchfield and A.W. Hubler. Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *arXiv:math.ST/0703715*, 2007.

[18] Z.-G. Yu, V. Anh and K.-S. Lau. Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome. *Physica*, **A 301**((1-4)):351–361, 2001.

[19] L.C. Zhao, C.C.Y. Dorea and C.R. Gonēalves. On determination of the order of a Markov chain. *Statistical Inference for Stochastic Processes*, **4**:273–282, 2001.