Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Study of Dependencies in the DNA of some Bacteria via Markov Models

C. Q. Silva

Relatório Técnico RTP-01/2000

Relatório Técnico
Série Pesquisa

Study of dependencies in the DNA of some bacteria via Markov models

Cibele Queiroz da Silva

March 20, 2000

Abstract

Dependencies in DNA sequencies are frequently modeled using Markov models. A possible approach is to treat the observed transition tables for each one of the orders to be tested as a contingency table and use log-linear models to test order and stationarity of the chain. One problem that arises in this process is that the null hypotheses to be tested are systematically rejected due to the large sample sizes that are necessary to use in such analyses. Those problems can be avoided by the use of the Bayesian Information Criterion, BIC, developed by Raftery (1986).

Key words: Markov models, log-linear models, BIC, DNA sequencies

Contents

1 Introduction

A large amount of DNA sequence data is now accessible through databases and various attempts have been made to use this information to model repeated patterns in such sequences. Among the possible models for studying those structures are Markov chains, which might describe the DNA sequence in terms of its nucleotide composition (i.e. as a string of letters from a four-letter alphabet, {A,C,G,T}).

In large DNA sequences as the genome of organisms (i.e. the total genetic information stored in the chromosomes of an organism) it is possible that variations among the transition probabilities along the genome occur. That implies the transition probabilities in a Markov chain are not stationary. In such case the structure of the process that generate the data is affected by translations of "time" axis (in our case, "time" means position).

In this work we fit Markov models and test stationarity for DNA sequences three bacteria: *Haemophilus influenzae*, *Mycoplasma genitalium* and *Methanococcus jannaschii* . The bacteria *Haemophilus influenzae* was once thought to be the causative agent for epidemic influenza. The complete nucleotide sequence of this bacteria is composed of 1,830,137 base pair. The *H. influenzae* genome sequence represents the only complete genome sequence from a free-living organism. The *Mycoplasma genitalium* genome is the smallest known genome of any free-living organism. Its complete nucleotide sequence is 580,070 base pairs. Comparison of this genome to that of *Haemophilus influenzae* suggests that differences in genome content are reflected as profound differences in physiology and metabolic capacity between these two organisms.

In the present work Markov models and stationarity tests were performed using Generalized Linear Models (GLM) together with the Bayesian Information Criterion (BIC). In section 2 we give some definitions and notations to be used through out the text. In section 3 we describe how to assess the order of a Markov chain and how to test stationarity using log-linear models. In section 4 we introduce the BIC and its use in hypothesis testing when large samples sizes are considered. In section 5 we analyse the data sets described above.

2 Estimation of transition probabilities

In this section we are going to give some general definitions and the needed background to the development of some further theory.

For a given subject, let X_t denote the response on a categorical variable at time t, $t = 0, 1, \cdots, T$. The sequence $(X_0, X_1, X_2, \cdots)$ is an example of a stochastic process, an indexed family of random variables.

A stochastic process is a rth-order Markov chain if, for all t, the conditional distribution of X_{t+1}, given $X_0, \cdots, X_t$, is identical to the conditional distribution of X_{t+1}, given $X_t, \cdots), X_{t-r+1}$. Given the states at the previous r times, the future behavior of the chain is independent of the past behavior before those r times. For a first order-Markov chain with

I possible states, the conditional probabilities $p_{ij}(t) = Pr[X_t|X_{t-1} = i]$ with $i, j = 1, \cdots, I$ are called *transition probabilities*. The extension for higher orders is immediate. If $p_{ij}(t)$ do not depend on t, the Markov chain is called homogeneous.

Statistical inference for Markov chain uses standard methods of categorical data analysis. To illustrate, we derive the Maximum Likelihood estimators of a first-order Markov chain.

Let $n_{ij}(t)$ denote the observed number of transitions from state i at time $t-1$ to state j at time t. For fixed t, $\{n_{ij}(t)\}$ form the two-way marginal table for dimensions $t-1$ and t of the full I^{T+1} contingency table (we are considering here the possibility of having several realizations of the stochastic process). For $n_{i+}(t)$ subjects classified in category i at time $t-1$, suppose the counts $\{n_{ij}(t), j = 1, \cdots, I\}$ have multinomial distribution with parameters $\{p_{ij}(t)\}$. Let $\{n_{i0} = n_{i+}(1)\}$ denote the initial counts, and suppose they also have a multinomial distribution with parameters $\{p_{i0}\}$. If all transitions are mutually independent, it follows from the factorization of the joint distribution of $X_0, \cdots, X_t$ under the first-order Markov chain that the Likelihood function is proportional to

$$\left(\prod_{i=1}^{I} p_{i0}^{n_{i0}}\right)\left(\prod_{t=1}^{T}\prod_{i=1}^{I}\left[\prod_{j=1}^{I} p_{ij}(t)^{n_{ij}(t)}\right]\right) \tag{1}$$

The Maximum Likelihood estimates for the parameters, considering $n = \sum_{i=1}^{I}\sum_{j=1}^{I} n_{ij}(t)$, are

$$\hat{p}_{i0} = \frac{n_{i0}}{n}$$

and

$$\hat{p}_{ij}(t) = \frac{n_{ij}(t)}{n_{i+}(t)}$$

For a Markov chain with stationary transition probability matrix, $p_{ij}(t) = p_{ij}$ for $t = 1, \cdots, T$, and the second factor in the Likelihood function (1) reduces to

$$\prod_{i=1}^{I}\prod_{j=1}^{I} p_{ij}(t)^{n_{ij}}$$

where $n_{ij} = \sum_{i=1}^{T} n_{ij}(t)$. The Maximum Likelihoodd estimates of the stationary transition probabilities are thus

$$\hat{p_{ij}} = \frac{n_{ij}}{n_{i+}(t)}$$

(Bishop et al 1975).

In the more general set up, when the transition probabilities are not necessarily stationary, we can represent the data on observed transitions in terms of the sufficient statistics $n_{ij}(t)$, and we can treat these statistics as if they were observed counts $n_{ijt} = n_{ij(t)}$ in an IxIxT contingency table, where the total for each level of the third variable is fixed equal to the total sample size n (see Anderson and Goodman - 1957).

When the transition probabilities are stationary, we can collapse the IxIxT table across the third variable, yielding an IxI table of counts $n_{ij} = n_{ij+}$ whose total equals $N = nT$.

Issues like testing stationarity and assessing the order of a Markov chain can be addressed via log-linear models. Those questions are discussed in the next section.

3 Markov chains and Generalized Linear Models

As we mentioned in the previous section, we can display the Markov chain data in contingency table form; the analyses of these data are formally equivalent to certain contingency table analyses based on log-linear models. A log-linear model is a linear model in the natural logarithms of the cell frequencies. A log-linear model takes part of a broad family of models called Generalized Linear Models. Those models are very flexible in the sense that the response variable can be modeled in such a way that the error component in the model does not need to have constant variance. In reality, it does not need to be even Gaussian. Response variables distributed as Binomial or Poisson are frequently modeled via GLM models.

Log-linear models are analogous to an analysis of variance (ANOVA) model. Considering a 4x4 contingency table, for $i, j = 1, 2, 3, 4$, we write

$$logm_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)}$$

In the model above m_{ij} represent the expected counts, u represents the overall mean, $u_{1(i)}$ and $u_{2(j)}$ represent the main effects of variables 1 and 2, respectively, and $u_{12(ij)}$ represent two-factor effect between variables 1 and 2. In this models some constraints must be observed

$$\sum_{i=1}^{4} u_{1(i)} = \sum_{j=1}^{4} u_{2(j)} = \sum_{i=1}^{4} u_{12(ij)} = \sum_{j=1}^{4} u_{12(ij)} = 0$$

To represent a Markov chain in terms of log-linear models, the number of variables needed equates the hypothesized order, r, of the chain plus one, i.e., $(r + 1)$.

3.1 Considerations about the response variable in a log-linear model

We mentioned in the previous sections that we can display the Markov chain data in contingency table form and then use standard methods of categorical data analysis to analyse the data. Besides that, if we consider that the counts in the contingency table have a Multinomial distribution, we can construct log-linear models using the GLM framework considering Poisson link function (for more details see McCullagh and Nelder 1989). The reasoning behind that is because when the total sample size is random, rather than fixed, the counts n_{ij} in the contingency table have Poisson distribution.

When the total sample size is fixed, the joint distribution of the counts n_{ij} is Multinomial. Birch (1963) showed that Maximum Likelihood estimates are the same for Multinomial sampling as for independent Poisson sampling. However, since most of the statistical software available for GLM models encorporate Poisson links, we developed the models to be presented in the further sections using this kind of analysis.

3.2 Assessing the order of a chain

Suppose we have a Markov chain which has either first or second order stationary transition probability matrix. If the Markov chain is really of first-order, then, for $i = 1, \cdots, I$, the transition probabilities are

$$p_{ijk} = p_{jk}$$

In the case of a second order Markov chain, the IxIxI table of counts n_{ijk} satisfies the general log-linear model

$$u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

and if we have a first-order chain, the table is described by the model

$$u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)}$$

Therefore, under the null hypothesis

$$\left\{ \; H_0 : \quad p_{ijk} = p_{jk} \right.$$

the terms which involve indexes 1 and 3, simultaneously, can be dropped from the saturated model and we obtain a reduced model that is good enough to explain the data. For higher orders, we proceed in the same way.

To test the reduced model (r) against the saturated model (s), we use the Maximum Likelihood estimators of m_{ijk} under both models and the Likelihood Ratio Test (LRT), that is

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{I} \sum_{k=1}^{I} \hat{m}_{ijk}^{(s)} log \left(\frac{\hat{m}_{ijk}^{(s)}}{\hat{m}_{ijk}^{(r)}} \right) \tag{2}$$

The quantity above is given by the residual deviance of the log-linear fit for the reduced model.

3.3 Stationary Transition Probabilities

As we mentioned earlier, for a first-order chain which has stationary transition probabilities, we can collapse the IxIxT table across the third variable, yielding an IxI table of counts

$n_{ij} = n_{ij+}$. In a first-order chain the transition probabilities are stationary if the expected frequencies for the table IxIxT satisfy the log-linear model in which for each *originating state*, the conditional distribution for one-step transitions from that state is identical at each of the T "times" for transitions. In order words, we need to verify whether or not the model which states that given the values of the "first state", the values of the "second state" are independent of the values of T is adequate for the data. This idea readily generalized for chain of higher orders. For more details see Agresti (1990) sections 5.3.3 and 5.3.4.

In the case of a first order Markov chain, to test stationarity is equivalent to the IxIxT table of counts n_{ijt} satisfy the model

$$u + u_{1(i)} + u_{2(j)} + u_{3(t)} + u_{12(ij)} + u_{13(it)}$$

Due to the large sample sizes that we are working with, null hypotheses of lower order Markov chains (and also null hypotheses of stationary transition probabilities) were systematically rejected in favor of higher order ones when we compared the quantity in equation (2) with its respective degrees-of-freedom to a Chi-square critical value. That means we need a different methodology to perform our tests. Such methodology was developed by Raftery (1986) and it is based in Bayesian theory. That is the issue we are going to treat in the next section.

4 BIC for hypothesis testing

According to Raftery (1995), with a large enough sample, the Likelihood Ratio Test (LRT) will detect any minor discrepancy between model and reality and reject even a good model. Raftery suggests instead a Bayesian approach to test hypothesis, particularly, using the Bayesian Information Criterion (BIC) approximation to the *Bayes Factor*. The Bayes Factor is a key quantity in the comparison of two competing models, say, M_1 and M_2, in a Bayesian framework for hypothesis testing.

The extent to which the data D support M_2 over M_1 is measured by the *posterior odds* for M_2 against M_1, i.e., the ratio of their *posterior probabilities*, that is,

$$\frac{P(M_2|D)}{P(M_1|D)} = \left[\frac{P(D|M_2)}{P(D|M_1)}\right]\left[\frac{P(M_2)}{P(M_1)}\right] \tag{3}$$

where $P(M_k)$ is the prior probability of the model M_k ($k = 1, 2$), so they incorporate prior beliefs the researcher has about data in study, $P(D|M_k)$ is the marginal probability of the data given M_k and $P(M_k|D)$ is the posterior probability of M_k given the data. The last probabilities are calculated using the Bayes' theorem.

The first factor on the right-hand side of equation (3) is called the Bayes Factor for M_2 against M_1, denoted by B_{21}.

According to Raftery (1994) when M_1 is nested within M_2, the Bayes Factor written on the scale of twice the logarithm (which defines the BIC) is approximated by

$$BIC_{21} = 2logB_{21} \approx G^2_{21} - d.f._{21}logn$$

where G^2_{21} is the deviance in the reduced model and $d.f._{21} = d_2 - d_1$ = number of degrees of freedom associated with the test.

Suppose M_2 represent a saturated model. The BIC value for such model is zero, and the saturated model is preferred to M_1 if $BIC_{21} > 0$, in which case M_1 can be considered not to fit the data well. When $BIC_{21} < 0$, M_1 is preferred to the saturated model, and the smaller BIC_{21} is (i.e. the more negative) the better the fit of M_1.

5 Analysis and Results

Using the methodology described in the previous sections, we analysed data sets of three bacteria, *Haemophilus Influenzae*, *Mycoplasma genitalium* and *Methanococcus Janannaschii.* Those data sets are composed of chunks of DNA randomly chosen along the each of fourth part of each data set. For each organism, four windows of size 50,000 nucleotides each were analysed. For each data set window it was evaluated the order of the Markov chain using log-linear models. In case the same order has been found for each one of the four data sets of an organism, a test to evaluate stationarity of the chain was performed. Such tests were also accomplished using log-linear models as described in section 3.3. As we mentioned in section 4, the sample sizes were taken into consideration in the performed tests by the use of the BIC.

5.1 Initial analyses

Figure 1 shows the probability distributions of the nucleotides for each one of the organisms in each of the four sampled "places" along the genome. Figure 1(d) shows the probability distributions of the nucleotides when the data corresponding to the four **places** of each organism are combined. The distributions between the organisms and also within them are roughly the same, with the genomes being A and T rich, with relative frequencies of at least about 30%. Such tendency is stronger for the genomes of the bacteria *Mycoplasma genitalium* and *Methanococcus Janannaschii.* For a first and second order chain, the more and the less frequent observed transitions are summarized in table 1. We observe that the more frequent transitions for both orders include either an A or a T type nucleotide in the last position of the string, while the less frequent transitions have either a C or a G type nucleotide at that position of the string. For a first order chain, the less frequent and the more frequent transitions that are common to the three organisms in study are {AA,TT} and TC, respectively. For a second order chain, the more frequent and the less frequent transitions that are common for the three organisms are {CCA,TAA} and {GTC,GAC}, respectively. Those results reflect the fact that the genome of those organisms are A and T rich. In table 1 we also observe that *Mycoplasma genitalium* and *Methanococcus jannaschii* share more patterns in common than those with *Haemophilus influenzae.* That might

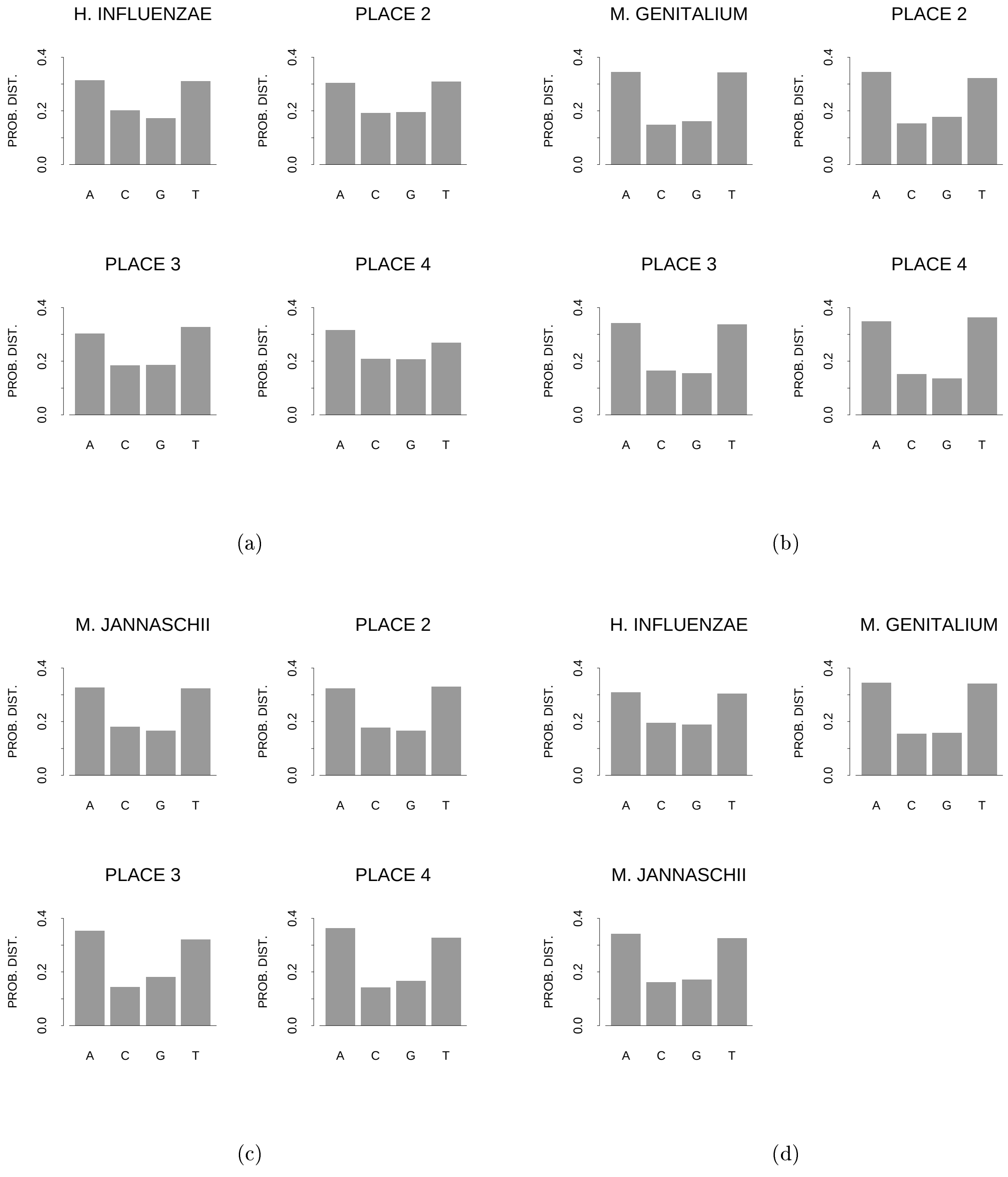

Figure 1: *PROBABILITY DISTRIBUTION FOR NUCLEOTIDES*

indicate that the former organisms have a closer genome content when compared to that of the *Haemophilus influenzae*.

Table 1: *Observed First and Second Order Transitions*

Transitions	Order	
	2	1
H. Influenzae		
More Freq. Trans.	CCA,CTT,TAA,TCA,TTT,GAA,AAA	TT,AA,CA
Less Freq. Trans.	GAG,GTC,GAC,CTC,AAC,TAG,AAG	AC,AG,TC
M. Genitalium		
More Freq. Trans.	GTT,CAA,GCT,CGA,CCA,TAA,TCA	AA,CA,TT
Less Freq. Trans.	TCG,CCG,ACG,GTC,GAC,GCC,CGC	CG,TC,AC
M. Jannnaschii		
More Freq. Trans.	GTT,AAA,TAA,AGA,TCT,CCA,ACA	AA,TT,CT
Less Freq. Trans.	TCG,ACG,GCG,CCG,GAC,GTC,CAC	TG,TC,GC

Table 2: *Haemophilus influenzae*

Place	Hypothesized order	d.f.	G^2	BIC
1	0	9	1774.79	1677.42
1	1	36	491.50	101.99
1	2	144	1006.04	-552.00
1	3	576	115.35	-5116.84
2	0	9	1085.73	988.36
2	1	36	705.82	316.31
2	2	144	715.75	-842.29
2	3	576	939.37	-5292.82
3	0	9	1446.56	1349.18
3	1	36	377.21	-12.30
3	2	144	857.75	-700.30
3	3	576	1134.67	-5097.52
4	0	9	2055.84	1958.46
4	1	36	584.30	194.79
4	2	144	1146.31	-411.74
4	3	576	1279.31	-4952.88

Table 3: *Mycoplasma genitaliun*

Place	Hypothesized order	d.f.	G^2	BIC
1	0	9	2412.25	2314.87
1	1	36	1117.36	727.84
1	2	144	772.65	-785.39
1	3	576	1056.80	-5175.39
2	0	9	2098.57	2001.19
2	1	36	1356.42	966.90
2	2	144	811.75	-746.29
2	3	576	1049.64	-5182.55
3	0	9	2352.41	2255.03
3	1	36	1512.05	1122.53
3	2	144	771.49	-786.55
3	3	576	988.18	-5244.01
4	0	9	2363.59	2266.21
4	1	36	798.58	409.06
4	2	144	798.22	-759.82
4	3	576	1031.9	-5200.22

Table 4: *Metahnococcus. jannaschii*

Place	Hypothesized order	d.f.	G^2	BIC
1	0	9	2051.63	1954.25
1	1	36	646.03	256.52
1	2	144	1607.06	49.01
1	3	576	1132.18	-5100.01
2	0	9	2195.53	2098.15
2	1	36	709.19	319.68
2	2	144	1341.73	-216.32
2	3	576	1158.67	-5073.52
3	0	9	2326.51	2229.13
3	1	36	885.45	495.93
3	2	144	1564.25	6.20
3	3	576	1134.22	-5097.97
4	0	9	1628.91	1531.53
4	1	36	935.08	545.57
4	2	144	1479.29	-78.75
4	3	576	1181.04	-5051.15

5.2 Testing order and stationarity of the chains

Tables 2 to 4 summarize the results of the fitted models for evaluating the order of a chain and whether or not it is stationary. In table 2 the first column indicates the piece of the data set we are analysing. Column 2 indicates the hypothesized order being tested with the respective degree-of-freedom of the test given in column 3. Column 4 indicates the Likelihood Ratio Test value and column 5 the corresponding BIC value. A negative value of BIC indicates that a chain whose hypothesized order is given in the left-hand side (in the same row) is preferred to a chain with higher order $(r+1)$. Therefore, BIC=-5116.85 indicates that the log-linear model for a third order chain is preferred to a saturated model for a fourth order chain. We observe that a second order chain seems to be adequate to describe the dependencies in the genome of *Haemophilus influenzae*. Tables 3 and 4 indicate that second and third order chains are adequate to describe those dependencies for the genomes of *Mycoplasma genitalium* and *Methanococcus jannaschii*, respectively.

The stationary tests resulted in the following BIC's for *Haemophilus influenzae*, Mycoplasma genitalium and *Methanococcus jannaschii*, respectively:

$$\begin{aligned} BIC &= 1291.6 - (144)log(200,000) = -466.104 \\ BIC &= 1382.3 - (144)log(200,000) = -375.374 \\ BIC &= 2117.8 - (576)log(200,000) = -4912.898 \end{aligned}$$

Those results indicate that, for the orders considered, each one of the chains is stationary.

6 Discussion

In this work we have fitted Markov models and tested stationarity for DNA sequnces of three bacteria: *Haemophilus influenzae*, *Mycoplasma genitalium* and *Methanococcus jannaschii*. Markov models and stationarity tests were performed using Generalized Linear Models (GLM) together with the Bayesian Information Criterion (BIC), due to the large sample sizes we had worked with. A second order chain was adequate to describe the dependencies in the genome of *Haemophilus influenzae* and *Mycoplasma genitalium*, while a third order chain was the adequated one for the data of *Methanococcus jannaschii*. The stationary tests showed that, for the orders considered, each one of the chains is stationary.

For future works the use of Hidden Markov Models might reveal some features that the present study was unable to do it.

7 References

Agresti, A. (1990). Categorical Data Analysis, 1st ed., John Wiley Sons, New York.

Anderson, T. W. and Goodman, L.A. (1957). Statistical Inference About Markov Chains. *Annals Math. Statist.*, **28**, 89-110.

Avery, P. J. (1987) The analysis of Intron data and their use in the detection of short signals. *Journal of Molecular Evolution*, **26**, pp. 335-340.

Bishop, Y. M. M., Fienberg, S. E., Holland, P. W., Mosteller, F. and Light, R. (1975). Discrete Multivariate Analysis: Theory and Practice, 1st ed., The MIT Press, Massachusetts.

McCullagh, P., Nelder, J.A. (1989) Generalized Linear Models, 2nd ed., Chapman Hall, London

Raftery, A., (1995). Bayesian Model Selection in Social Research (with discussion) - *Sociological Methodology* (Peter V. Marsden, Editor).

Raftery, A., Tavare, S. (1994). Estimation and Modelling Repeated Patterns in High Order Markov Chains with the Mixture Transition Distribution Model, *Applied Statistics*, **43**, No. 1, pp. 179-199.

Raftery, A., (1986). Choosing Models for Cross-classifications, *Amer. Sociol. Rev.* **51:** pp. 145-146.