Models Applied to DNA Sequences with Multinomial Correlated Responses

Beatriz C. D. Cuyabano[1], Hildete P. Pinheiro[1] and Aluísio Pinheiro[1]
[1] *Department of Statistics, University of Campinas, Brazil*

Abstract

Multinomial multivariate models are proposed to describe the codon frequencies in DNA sequences, as well as the order and frequency that nucleotide bases have in each codon considering the dependence among the bases inside a codon. Logistic regressive models are used with different dependence structures on the three codon positions. Also, multinomial extensions of the Bahadur's representation are proposed to model correlated multinomial data. An application of these models to the NADH4 gene from human mitochondrial genome is presented. AIC, BIC and the leave-one-out cross validation are employed to compare the various models peformance.

Keywords and phrases: Multinomial correlated data; Generalized linear models; Statistical genetics; DNA sequences.

1 Introduction

Genetic sequencing has become a very important field of research all over the world, and genomic studies increase everyday. This fact is easily comprehensible, since the amount of information one can obtain by analyzing a single gene[1] is very large, and it is able to provide answers to many questions. Besides, improvements on computer technology, accessibility to genomic data, and the increasing research around this theme favor the studies and the improvement of techniques to DNA analysis.

This study focus on the analysis of codon[2] frequencies from the NADH4 gene from the human genome, and also its sequence of nitrogen bases - Thymine (T), Cytosine (C), Adenine (A) and Guanine (G) based on the *Cambridge Reference Sequence - CRS* (Anderson et al., 1981; Andrews et al., 1999).

The importance of the statistical studies involving this subject is specially due to the interest that researchers have in comprehending and predicting genetic structures. The complexity of DNA demands models with high levels of

[1] DNA segment containing the necessary information for protein synthesis

[2] triple of adjacent nitrogen bases

details, as in the logistic regressive models (Bonney et al., 1994), which analyze the dependence among the three positions in a codon, by using the previous positions as covariates for logistic models of the next ones.

We employ the Bahadur's representation (Bahadur, 1961), that defines a joint probability function for correlated binary data. Cox (1972) presents this representation as an alternative to logistic models when writting the joint probability of correlated binary data, directly in terms of the correlation parameters, instead of using log-probabilities. Also, Zhao e Prentice (1990) use the Bahadur's representation as a particular case of the exponential quadratic model to correlated binary data.

The Bahadur's representation has also been used in studies of binary responses to individuals separated in groups (Stefanescu e Turnbull, 2003), and it is found in the literature for studies of longitudinal data (Fitzmaurice et al., 1993; Fitzmaurice, 1995); Parzen et al. (2009) uses this representation in time series data with a general autocorrelation approach and these models have estimation methods that involve autoregressive models.

The Bahadur's representation was also used for analyzing mutations in genetic structure (Pinheiro et al., 1999). Therefore, the use of this representation, expanded to multinomial correlated data, is pertinent for modeling dependence structures inside a codon.

The logistic models when first published (Bonney, 1986) were applied to model family data; later, Bonney et al. (1994) presented a genetic approach of these models, applying them to a single DNA sequence, and the Bahadur's representation, although suggested as possible and briefly described by Bahadur, has not been used to model multivariate correlated data in the field of genetics.

In section 2 the probabilistic model is presented, as well as possible covariates that may be used in the context. The logistic regressive models are discussed in section 3 and the models based on the Bahadur's representation are presented in section 4. The maximum likelihood estimation (MLE) procedure is discussed in section 5. These models are applied to a real data set on section 6. A discussion follows in section 7.

2 Multinomial Models for Codons Probabilities

Considering the codon as a triplet of adjacent nitrogen bases represented by the vector $\mathbf{Y}_i = (Y_{1i}, Y_{2i}, Y_{3i})$, where Y_{ki} can assume the bases T, C, A or G, there are 64 possible distinct codons. Since four of them are stop codons (thus, non-effective), only 60 codons are used in the analysis being considered independent from one another, with a multinomial distribution $(\mathbf{Y}_1, ..., \mathbf{Y}_{60}) \sim M\,(p_1, ..., p_{60})$, and at a sample of N codons:

$$P\left([\#\mathbf{Y}_i = n_i]_{i=1}^{60}\right) \quad = \quad \frac{N!}{\prod_{i=1}^{60} n_i!} \prod_{i=1}^{60} p_i^{n_i}. \qquad (1)$$

The observed probabilities from this sample are then the maximum likelihood estimates of the multinomial distribution, $\hat{p}_i = n_i/N$, for all $i = 1, ..., 60$.

A codon determines one of the 20 existing amino acids. More than one codon can determine the same amino acid, and these are called synonymous codons. The stop codons, previously mentioned, don't synthesize any amino acid, and are TAA, TAG, AGA and AGG. The vector $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i})$ represents the covariates associated to each codon, and are the same used by Bonney (1994). These covariates are the following: AARISK, that measures the risk of mutation for an amino acid, AVDIST, a measure of how typical an amino acid is, and TSCORE that measures the number of single changes in only one nitrogen base, that may cause a codon to become a stop codon (e.g., codon TTA has TSCORE 1, for only changing the second T to A that a stop codon is obtained, and codon TGA has TSCORE 2, for when changing G to A or T to A a stop codon is obtained). Synonymous codon don't necessarily have the same values of AARISK or TSCORE. It is important to note that all three covariates are not associated to the stop codons.

AARISK and AVDIST are calculated based on the following chemical properties of the amino acids: composition (c), polarity (p) and molecular volume (v); also based on α, β and γ which represents the inverse mean squares, respectively for the composition, polarity and molecular volume of all the 20 amino acids (Granthan, 1974). AARISK is the weighted mean of the distance between an amino acid and the others, and AVDIST is the non weighted mean of the distance between an amino acid and the others, and the smallest the values of AVDIST, the most typical the amino acid is. The distances to obtain these covariates are given by,

$$D_{ij} = \sqrt{\left[\alpha\left(c_i - c_j\right)^2 + \beta\left(p_i - p_j\right)^2 + \gamma\left(v_i - v_j\right)^2\right]}. \tag{2}$$

Since $\sum_{i=1}^{60} P(\mathbf{Y}_i|\mathbf{X}_i) = 1$, it is possible to define the following weighted probability $WP(\mathbf{Y}_i|\mathbf{X}_i)$ that will ensure correct estimates when modeling,

$$WP(\mathbf{Y}_i|\mathbf{X}_i) = \frac{P(\mathbf{Y}_i|\mathbf{X}_i)}{\sum_{i=1}^{60} P(\mathbf{Y}_i|\mathbf{X}_i)}. \tag{3}$$

Hence, substituting p_i by $WP(\mathbf{Y}_i|\mathbf{X}_i)$, equation (1) is now written as,

$$P\left([\#\mathbf{Y}_i = n_i]_{i=1}^{60}\right) = \frac{N!}{\prod_{i=1}^{60} n_i!} \prod_{i=1}^{60} \left[\frac{P(\mathbf{Y}_i|\mathbf{X}_i)}{\sum_{i=1}^{60} P(\mathbf{Y}_i|\mathbf{X}_i)}\right]^{n_i}. \tag{4}$$

3 Logistic Regressive Models

The logistic regressive models introduce the dependence structure by using the previous positions in the codon as covariates to model the next. Basically, to model the second position, the first one becomes a covariate, and to model the third position, the first and the second ones become covariates. Thus, we are modeling the variables conditionally to the previous ones, and using the multiplication theorem to write the probability of observing codon as the product of these conditional probabilities,

$$P(\mathbf{Y}_i|\mathbf{X}_i) = P(Y_{1i}|\mathbf{X}_i)P(Y_{2i}|Y_{1i},\mathbf{X}_i)P(Y_{3i}|Y_{1i},Y_{2i},\mathbf{X}_i). \tag{5}$$

As explained before, each of the three positions inside the codon assume one of the bases T, C, A or G, therefore, conditionally to the previous positions, Y_{ki} has a multinomial distribution $(Y_{ki}|Y_{1i},...,Y_{k-1,i},\mathbf{X}_i) \sim M(\pi_{k0i},\pi_{k1i},\pi_{k2i},\pi_{k3i})$, where $\pi_{kji} = P(Y_{ki} = j|Y_{1i},...,Y_{k-1,i},\mathbf{X}_i)$.

In order to simplify notation, we codify numerically the nitrogen bases as $T = 0$, $C = 1$, $A = 2$ and $G = 3$. Now, variables Z_{kji} $(k, j = 1, 2, 3)$ are created, such that $Z_{kji} = 1$ if $Y_{ki} = j$, and $Z_{kji} = 0$ otherwise. Table 1 displays the relationship between these variables, the nitrogen bases and the numerical codification.

Table 1: Codification of Nitrogen Bases

Base	Numerical Code	$(Z_{k1i}, Z_{k2i}, Z_{k3i})$
T	0	$(0,0,0)$
C	1	$(1,0,0)$
A	2	$(0,1,0)$
G	3	$(0,0,1)$

The multinomial logit function is used to work with the multinomial response variables and the covariates for modeling the data, with the probability of Y_{ki} being the base T (or 0, numerically) as the reference. The logit is then written as $\theta_{kji} = \log(\pi_{kji}/\pi_{k0i})$ for every $j = 1, 2, 3$; consequently $\pi_{kji} = e^{\theta_{kji}}/(1 + \sum_{i=1}^{3} e^{\theta_{kji}})$ for every $j = 1, 2, 3$ and $\pi_{k0i} = 1/(1 + \sum_{i=1}^{3} e^{\theta_{kji}})$.

The distribution of the $k-th$ position in a codon is then given by

$$P(Y_{ki}|Y_{1i},...,Y_{k-1,i},\mathbf{X}_i) = \pi_{k1i}^{z_{k1i}}\pi_{k2i}^{z_{k2i}}\pi_{k3i}^{z_{k3i}}\pi_{k0i}^{1-\sum_{j=1}^{3} z_{kji}}, \tag{6}$$

substituting in equation (5) and writing π_{kji} as a function of the logit θ_{kji},

$$P(\mathbf{Y}_i|\mathbf{X}_i) = \prod_{k=1}^{3} \frac{e^{\sum_{j=1}^{3} z_{kji}\theta_{kji}}}{(1 + \sum_{j=1}^{3} e^{\theta_{kji}})}, \tag{7}$$

and from equation (4) we derive the log-likelihood,

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^{3} \left[\sum_{j=1}^{3} z_{kji}\theta_{kji} - \log(1 + \sum_{j=1}^{3} e^{\theta_{kji}}) \right] \\ &\quad - \sum_{i=1}^{60} n_i \log \left[\sum_{i=1}^{30} \prod_{k=1}^{3} \frac{e^{\sum_{j=1}^{3} z_{kji}\theta_{kji}}}{(1 + \sum_{j=1}^{3} e^{\theta_{kji}})} \right]. \end{aligned} \tag{8}$$

Independent Model

The independent model, as intuitively, assumes no dependence among the positions inside a codon. So, each position's logit is a function only of the covariates, having a total of 12 parameters. Then, for every $k, j = 1, 2, 3$, the logits are given by

$$\theta_{kji} = \alpha_{kj} + \sum_{p=1}^{3} \beta_p X_{pi}. \tag{9}$$

Equally Predictive Model

This model considers dependence among the positions in a codon, but assumes that the influence of a position on the next ones is always the same and also equal to the influence of every other previous positions, having a total of 15 parameters. Then, for every $j = 1, 2, 3$, the logits are given by

$$\theta_{1ji} = \alpha_{1j} + \sum_{p=1}^{3} \beta_p X_{pi}, \tag{10}$$

$$\theta_{2ji} = \alpha_{2j} + \sum_{s=1}^{3} \gamma_s Z_{1si} + \sum_{p=1}^{3} \beta_p X_{pi} \text{ and} \tag{11}$$

$$\theta_{3ji} = \alpha_{3j} + \sum_{s=1}^{3} \gamma_s (Z_{1si} + Z_{2si}) + \sum_{p=1}^{3} \beta_p X_{pi}. \tag{12}$$

First Order Markov Structure

The first order markov structure considers influence on a codon's position only by the immediate previous one, having a total of 18 parameters. Then, for every $j = 1, 2, 3$, the logits are given by

$$\theta_{1ji} = \alpha_{1j} + \sum_{p=1}^{3} \beta_p X_{pi}, \tag{13}$$

$$\theta_{2ji} = \alpha_{2j} + \sum_{s=1}^{3} \gamma_{1,s} Z_{1,si} + \sum_{p=1}^{3} \beta_p X_{pi} \text{ and} \tag{14}$$

$$\theta_{3ji} = \alpha_{3j} + \sum_{s=1}^{3} \gamma_{2,s} Z_{2,si} + \sum_{p=1}^{3} \beta_p X_{pi}. \tag{15}$$

Additive Model

The additive model is the most complete model proposed. It considers the influence on a codon's position by every previous ones, having a total of 21 parameters. Then, for every $j = 1, 2, 3$, the logits are given by

$$\begin{aligned}
\theta_{1ji} &= \alpha_{1j} + \sum_{p=1}^{3} \beta_p X_{pi}, && (16)\\
\theta_{2ji} &= \alpha_{2j} + \sum_{s=1}^{3} \gamma_{1s} Z_{1si} + \sum_{p=1}^{3} \beta_p X_{pi} \text{ and} && (17)\\
\theta_{3ji} &= \alpha_{3j} + \sum_{s=1}^{3} (\gamma_{2s} Z_{1si} + \gamma_{3s} Z_{1si}) + \sum_{p=1}^{3} \beta_p X_{pi}. && (18)
\end{aligned}$$

4 Model Based on the Bahadur's Representation

The Bahadur's representation was first introduced to describe correlated binary data. Consider a set $Y_1, Y_2, ..., Y_K$ in which every $Y_k \sim ber(\pi_k)$, for every $k = 1, 2, ..., K$. If the variables are independent, their joint distribution is given by,

$$P(\mathbf{Y} = \mathbf{y}) = \prod_{k=1}^{K} \pi_k^{y_k} (1 - \pi_k)^{1-y_k}. \tag{19}$$

Creating the set $U_1, U_2, ..., U_K$ of standardized variables, such that $U_k = (Y_k - \pi_k)/\sqrt{\pi_k(1 - \pi_k)}$, considering the measures of correlation $\rho_{kl} = E(U_k U_l)$, $\rho_{kls} = E(U_k U_l U_s)$, ... , $\rho_{12...K} = E(U_1 U_2 ... U_K)$, and knowing that this representation does not assume independence, the joint distribution of the variables shall be,

$$P(\mathbf{Y} = \mathbf{y}) = \left[\prod_{k=1}^{K} \pi_k^{y_k} (1 - \pi_k)^{1-y_k} \right] f(\boldsymbol{\rho}, \mathbf{u}), \tag{20}$$

where $f(\boldsymbol{\rho}, \mathbf{u})$ introduces a dependence structure as follows,

$$\begin{aligned}
f(\boldsymbol{\rho}, \mathbf{u}) &= 1 + \sum_{k>l} \rho_{kl} u_k u_l + \sum_{k>l>s} \rho_{kls} u_k u_l u_s \\
&\quad + ... + \rho_{12...K} u_1 u_2 ... u_K. && (21)
\end{aligned}$$

Returning to the multinomial case, each codon is a set $\mathbf{Y}_i = (Y_{1i}, Y_{2i}, Y_{3i})$. Thus, assuming independence and considering the covariates, $(Y_{ki}|\mathbf{X}_i) \sim M(\pi_{k0i}, \pi_{k1i}, \pi_{k2i}, \pi_{k3i})$, where $\pi_{kji} = e^{\theta_{kji}}/(1 + \sum_{i=1}^{3} e^{\theta_{kji}})$ for every $j = 1, 2, 3$ and $\pi_{k0i} = 1/(1 + \sum_{i=1}^{3} e^{\theta_{kji}})$. Given the independence assumption, we use the first regressive model described, in which $\theta_{kji} = \alpha_{kji} + \sum_{p=1}^{3} \beta_p X_{pi}$. Hence, if the positions in a codon are independent, their joint distribution is given by equation (7).

The set of standardized variables $U_{k1i}, U_{k2i}, U_{k3i}$ for every $k = 1, 2, 3$ is now created, such that $U_{kji} = (Z_{kji} - \pi_{kji})/\sqrt{\pi_{kji}(1 - \pi_{kji})}$. In the multinomial extension of the Bahadur's representation, we shall consider only the second order correlation measures, in order to reduce the number of parameters. Besides, four possible extensions will be presented, with different dependence structures $f(\boldsymbol{\rho}, \mathbf{u})$, in order of complexity (number of parameters). Finally, the joint distribution of the multinomial variables that represent the nitrogen bases at the positions in a codon is given by

$$P(\mathbf{Y}_i|\mathbf{X}_i) = \left[\prod_{k=1}^{3} \frac{e^{\sum_{j=1}^{3} z_{kji}\theta_{kji}}}{(1 + \sum_{j=1}^{3} e^{\theta_{kji}})}\right] f(\boldsymbol{\rho}, \mathbf{u}_i), \tag{22}$$

and from equation (4) we derive the log-likelihood,

$$\begin{aligned} \ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^{3} \left[\sum_{j=1}^{3} z_{kji}\theta_{kji} - \log(1 + \sum_{j=1}^{3} e^{\theta_{kji}})\right] \\ &\quad + \sum_{i=1}^{60} n_i \log\left[f(\boldsymbol{\rho}, \mathbf{u}_i)\right] \\ &\quad - \sum_{i=1}^{60} n_i \log\left[\sum_{i=1}^{60} P(\mathbf{Y}_i|\mathbf{X}_i)\right], \end{aligned} \tag{23}$$

where $P(\mathbf{Y}_i|\mathbf{X}_i)$ is given by equation (22).

It is very important to make sure that $P(\mathbf{Y}_i|\mathbf{X}_i)$ is greater than zero and smaller than one, since it defines a probability measure. The logistic part of the model already defines the probabilities π_{kji} strictly positive and the weighted probabilities ensures that they are smaller than one. Therefore the following constraint must be obeyed,

C1. $f(\boldsymbol{\rho}, \mathbf{U}_i) > 0, \quad \forall \ i = 1, ..., 60.$

In addition, as the vector of parameters $\boldsymbol{\rho}$ defines correlation measures, a second constraint is required,

C2. $\rho_* \in [-1, 1], \quad \forall \ \rho_* \in \boldsymbol{\rho}.$

Location Dependent Model

This model considers only the change from one position to another, without taking into account the nitrogen bases. In other words, it assumes, for example, that a transition from an Adenine at the first position, to a Cytosine at the second, is equally correlated as a transition from a Thymine at the first position, to a Guanine at the second. Hence, $f(\boldsymbol{\rho}, \mathbf{u}_i)$ is given by

$$\begin{aligned} f(\boldsymbol{\rho}, \mathbf{u}_i) &= 1 + \sum_{j,s} (\rho_{12} u_{1ji} u_{2si} + \rho_{13} u_{1ji} u_{3si}) \\ &\quad + \sum_{j,s} \rho_{23} u_{2ji} u_{3si}, \end{aligned} \tag{24}$$

for $j, s = 1, 2, 3$, such that $\rho_{12} = E(U_{1ji} U_{2si})$, $\rho_{13} = E(U_{1ji} U_{3si})$ and $\rho_{23} = E(U_{2ji} U_{3si})$ for every $j, s = 1, 2, 3$. Counting these 3 correlation parameters $\boldsymbol{\rho}$ along with the 9 intercepts $\boldsymbol{\alpha}$ and the 3 covariates' parameters $\boldsymbol{\beta}$ from the logit, this model has a total of 15 parameters.

Transition Dependent Model

This model considers only the change from a nitrogen base to another, regardless of the positions. In other words, it assumes, for example, that a transition from an Adenine at the first position, to a Thymine at the second, is equally correlated as a transition from an Adenine at the second position, to a Thymine at the third. Hence, $f(\boldsymbol{\rho}, \mathbf{u}_i)$ is given by

$$\begin{aligned} f(\boldsymbol{\rho}, \mathbf{u}_i) &= 1 + \sum_{j,s} \rho_{js} (u_{1ji} u_{2si} + u_{1ji} u_{3si}) \\ &\quad + \sum_{j,s} \rho_{js} u_{2ji} u_{3si}, \end{aligned} \tag{25}$$

for $j, s = 1, 2, 3$, such that $\rho_{js} = E(U_{kji} U_{lsi})$ for every $j, s = 1, 2, 3$ and $k \neq l$. Counting these 9 correlation parameters $\boldsymbol{\rho}$ along with the 9 intercepts $\boldsymbol{\alpha}$ and the 3 covariates' parameters $\boldsymbol{\beta}$ from the logit, this model has a total of 21 parameters.

Semi-Location and Transition Dependent Model

This model considers the change from a nitrogen base to another, but also takes into account the gap of this change. In other words, it assumes, for example, that a transition from the first position to the second and a transition from the second position to the third are equally correlated, but differently from a transition from the first position to the third. Hence, $f(\boldsymbol{\rho}, \mathbf{u}_i)$ is given by

$$\begin{aligned} f(\boldsymbol{\rho}, \mathbf{u}_i) &= 1 + \sum_{j,s} \rho_{1,js} (u_{1ji} u_{2si} + u_{2ji} u_{3si}) \\ &\quad + \sum_{j,s} \rho_{2,js} u_{1ji} u_{3si}, \end{aligned} \tag{26}$$

for $j, s = 1, 2, 3$, such that $\rho_{1,js} = E(U_{1ji} U_{2si}) = E(U_{2ji} U_{3si})$ and $\rho_{2,js} = E(U_{1ji} U_{3si})$ for every $j, s = 1, 2, 3$. Counting these 18 correlation parameters $\boldsymbol{\rho}$

along with the 9 intercepts $\boldsymbol{\alpha}$ and the 3 covariates' parameters $\boldsymbol{\beta}$ from the logit, this model has a total of 30 parameters.

Location and Transition Dependent Model

This model is considered the *full dependence model*, since it takes into account the nitrogen bases and the positions that are changing. Hence, $f(\boldsymbol{\rho}, \mathbf{u}_i)$ is given by

$$\begin{aligned} f(\boldsymbol{\rho}, \mathbf{u}_i) &= 1 + \sum_{j,s} (\rho_{1j,2s} u_{1ji} u_{2si} + \rho_{1j,3s} u_{1ji} u_{3si}) \\ &\quad + \sum_{j,s} \rho_{2j,3s} u_{2ji} u_{3si}, \end{aligned} \tag{27}$$

for $j, s = 1, 2, 3$, such that $\rho_{1j,2s} = E(U_{1ji}U_{2si})$, $\rho_{1j,3s} = E(U_{1ji}U_{3si})$ and $\rho_{2j,3s} = E(U_{2ji}U_{3si})$. Counting these 27 correlation parameters $\boldsymbol{\rho}$ along with the 9 intercepts $\boldsymbol{\alpha}$ and the 3 covariates' parameters $\boldsymbol{\beta}$ from the logit, this model has a total of 39 parameters.

5 Parameters Estimation

The estimation of the parameters for the regressive models is relatively an easy task. The logistic function ensures that the probabilities estimated are always greater than zero, and there are no restrictions upon the $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta})$. Some difficulty appears, however, when estimating the models based on the Bahadur's representation.

Maximizing the log-likelihood from equation (23) subject to both constraints C1 and C2 is a computationally complex problem; therefore, instead of estimating the full set of parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$ simultaneously, we will do this in two steps. The main problem that appears when trying to maximize the log-likelihood to obtain all the parameters of interest at once, is related to the first constraint. Thus, it is hard to manipulate optimization routines subject to non-linear box-constraints.

The two-step method consists of first obtaining the estimates for $(\boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{\rho} = \mathbf{0})$, which is the independent regressive model. Secondly, we obtain the estimates for $(\boldsymbol{\rho} | \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$, reducing the first constraint to a linear one and making it easy to handle with optimization routines subject to linear box-constraints. The log-likelihood for the second step, since $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ and $\hat{\pi}_{kji}$ are known estimated values, is then given by

$$\begin{aligned} \ell(\boldsymbol{\rho}) &= \sum_{i=1}^{60} n_i log \left[f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) \right] \\ &\quad - \sum_{i=1}^{60} n_i \log \left[\sum_{i=1}^{60} P(\mathbf{Y}_i | \mathbf{X}_i) \right], \end{aligned} \tag{28}$$

where $P(\mathbf{Y}_i|\mathbf{X}_i)$ is given by equation (22).

6 Application

The proposed models were applied to 30 sequences of the NADH4 gene, among them, the CRS, obtained at http://www.ncbi.nlm.nhi.gov. To compare the models, three measures of adjustment were verified: the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Square Sum of the Errors (SSE). The first two take into account the log-likelihood and the number of parameters, and the third, only the distance between observed and fitted values. Values obtained for the regressive models and those based on the Bahadur's representation are shown in Tables 2 and 3, respectively.

Table 2: Results for the Regressive Models

Model	AIC	BIC	SSE
Independent	101723.21	101813.57	0.0052
Eq. Predictive	100939.08	101052.03	0.0037
First Order Markov	99766.62	99902.16	0.0019
Additive	**99596.07**	**99754.21**	**0.0017**

Table 3: Results of Models Based on the Bahadur's Representation

Model	AIC	BIC	SSE
Location	101588.80	101701.75	0.0048
Transition	100740.39	100898.52	0.0034
Semi-Location and Transition	2***100030.35**	2***100256.26**	2***0.0024**
Location and Transition	100144.70	100438.38	0.0027

Among the regressive models, the Additive Model is the one that presents better fit, with smallest AIC, BIC and SSE, compared to the others. Among the models based on the Bahadur's representation, the Semi-location and Transition Dependent Model is the best one. If we compare these models, the Additive is definitely the best one.

Figures 1 and 2 show the fitted versus the observed probabilities for each of the 60 codons, for the Additive and Semi-location and Transition Models, respectively. It is possible to see that the codons with smaller probabilities are the ones that have a better adjusted values for both models. In addition, while the Additive Model fits better the probabilities of some codons - codon number 3 (ATT), for example - it does fit worse the probabilities than other codons - codon number 23 (ACC), for example. Although the AIC, BIC and SSE make us choose the Additive Model, in a general look at the graphics however, it is hard to define which model has a better fit for the data.

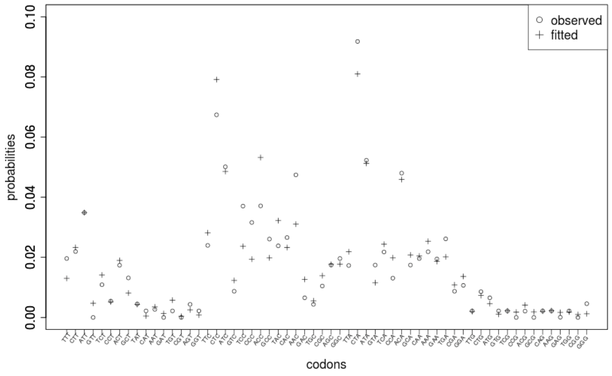

Figure 1: Observed versus Fitted Values of the Additive Model

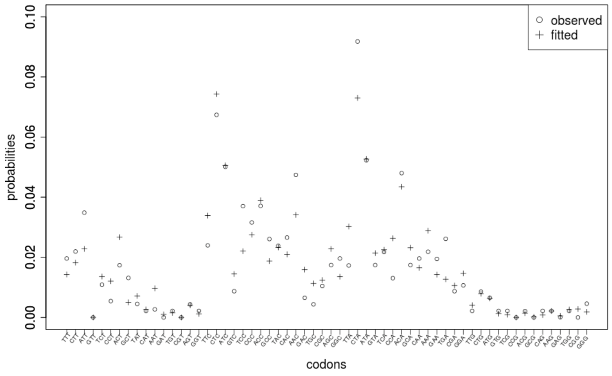

Figure 2: Observed versus Fitted Values of the Semi-Location and Transition Model

Another method to evaluate model fitness is cross-validation. The technique used here is the *leave-one-out*, in which a single item from the sample is held out, then, the model is adjusted for all the remaining others. Next, the obtained fitted value is compared with the item that was not used for modeling. This procedure is repeated for each item of the sample, and the SSE for each comparison is analyzed, mainly observing its mean value and variance. A disadvantage of this procedure is that since it is repeated for each item of the sample, it can be a computationally intensive procedure.

We applied the cross-validation to both, the Additive and Semi-location and Transition Model, and the Mean Square Error (MSE) and variance of the SSE for each comparison are displayed in Table 4, where we can see that both are smaller for the Additive Model.

Table 4: Cross-Validation Results

Model	MSE	$Var(SSE)$
Additive	0.0017	1.45×10^{-12}
Semi-Location and Transition	0.0024	8.06×10^{-10}

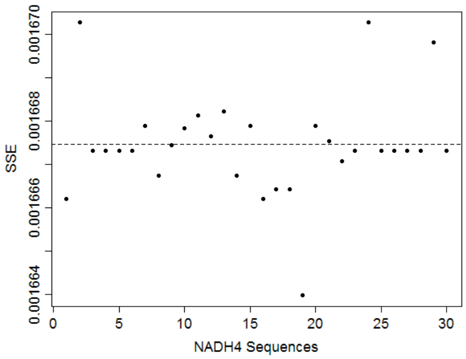

Figure 3: Cross-Validation's SSE of the Additive Model

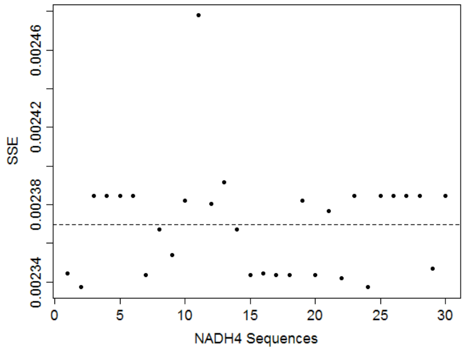

Figure 4: Cross-Validation's SSE of the Semi-Location and Transition Model

It is possible to see in figures 3 and 4, where the dotted line represents the mean values for the SSE obtained with the cross-validation, that the additive model is indeed the one with smaller SSE variance. On the other hand, it is possible to see that the additive model apparently produces a few outliers, while the other produces only one.

7 Discussion

On a first glance at the results obtained, not only the AIC, BIC and SSE indicate that the additive model is more suitable for the DNA data, but also the cross-validation is consistent with these results, presenting smaller MSE and less variation to the SSE.

There are some interesting points, though, to note about the adjustment of the models and the decision about which model should be used to fit this sort of DNA data. The first one is that the difference between the AIC from the Additive Model and the Semi-location and Transition Model is 434.28, which represents 0.4360% of the Additive Model's AIC. The difference in the BIC is $502,05$, which represents $0,5033\%$ of the Additive Model's BIC (in other words, the Semi-location and Transition Model's AIC and BIC are respectively 1.0044 and 1.0050 times the values obtained for the Additive's). Hence, in terms of proportion, the values obtained for the best model based on the Bahadur's representation are not really much greater than those obtained for the best regressive model.

Another important matter in these models is the number of parameters. When looking only to codons, the number of parameters for the Additive Model is indeed smaller than that for the Semi-location and Transition Model. If we decide, however, to expand it for more than just these 3 positions in a codon, there will be more parameters in the Additive Model. For both models, if there are K positions to be considered as dependent, there will be $3K$ intercept parameters and 3 covariates parameters. The difference is in the dependence parameters. For the additive model, there will be $3K(K-1)/2$ parameters, and for the Semi-location and Transition Model, there will be $9(K-1)$. Table 5 displays the increase in the number of dependence parameters for these two models, as we add positions.

Table 5: Number of Parameters for Additive Model and Semi-Location and Transition Model

	Total of Dependent Positions								
Model	1	2	3	4	5	6	7	...	100
Additive	0	3	9	18	30	45	63	...	14850
Semi-Location and Trans.	0	9	18	27	36	45	54	...	891

One might ask why not consider the first order Markov structure, that has smaller AIC, BIC and SSE than the Semi-location and Transition Model, and for K positions, will have $3(K-1)$ dependence parameters, less than the Semi-location and Transition Model. The problem with this model is exactly the fact that it is a first order Markov structure. It does not look backwards to the whole sequence, as does the additive or the semi-location & transition models, but only to the immediately previous position.

In summary, it seems that probably for modeling the codons probabilities, the Additive Model has a better fit. However, as we expand the model for greater dependence chains, or maybe even for the whole gene, as shown in Table 5, the number of parameters increases greatly, and due to the fact that proportionally, the AIC and BIC of the Semi-location and Transition Model are not much greater than those of the additive model, maybe the model based on the Bahadur's representation should fit better, as the number of parameters is far much smaller.

References

Anderson, S.; Bankier, A. and Barrell, B. (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290** 457–465

Andrews, R.; Kubacka, I.; Chinnery, P.; Lightowlers, R.; Turnbull, D. and Howell, N. (1999). Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. *Nature Genet.* **23** 147

Bahadur, R.R. (1961). A Representation of the Joint Distribution of Responses to n Dichotomous Items. *In studies in Item Analysis and Prediction* 158–176 H. Solomon (ed.). Stanford University Press

Bonney, G.E. (1986). Regressive Logistic Models for Familial Disease and Other Binary Traits. *Biometrics* **42** 611–625

Bonney, G.E.; Amfoh, K. and Shaw, R. (1994). The Use of Logistic Models for the Analysis of Codon Frequencies of DNA Sequences in Terms of Explanatory Variables. *Biometrics* **50** 1054–1063

Cox, D.R. (1972). The Analysis of Multivariate Binary Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **Vol. 21, No. 2** 113–120

Fitmaurice, G.M.; Laird, N.M. and Rotnitzky, A.G. (1993). Regression Models for Discrete Longitudinal Responses. *Statistical Science* **Vol. 8, No. 3** 284–299

Fitzmaurice, G.M. (1995). A Caveat Concerning Independence Estimating Equations with Multivariate Binary Data. *Biometrics* **Vol. 51, No. 1** 309–317

Granthan, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **185** 862–864

Lipsitz, S.R.; Parzen, M.; Ghosh, S.; Sinha, D.; Fitzmaurice, G.M.; Ibrahim, J.G. and Mallick, B.K. (2011). A Generalized Linear Mixed Model for Longitudinal Binary Data with a Marginal Logit Link Function. *accepted for publication in the Annals of Applied Statistics*

Pinheiro, H.P.; Seillier-Moiseiwitsch, F. and Sen, P.K. (1999). Modeling the Mutation Process in the HIV Genome. *Research Report #13/1999. University of Campinas, Brazil*

Stefanescu, C. and Turnbull, B.W. (2003). Likelihood Inference for Exchangeable Binary Data with Varying Cluster Sizes. *Biometrics* **Vol. 59, No. 1** 18–24

Zhao, L.P. and Prentice, R.L (1990). Correlated Binary Regression Using a Quadratic Exponential Model. *Biometrika* **Vol. 77, No. 3** 642–648