STATISTICS,OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **1**, December 2013, pp 8–28.
Published online in International Academic Press (www.IAPress.org)

Mixture models of geometric distributions in genomic analysis of inter-nucleotide distances

Adelaide Freitas[1, *], Vera Afreixo[1] and Sara Escudeiro[2]

[1] *Department of Mathematics & CIDMA, University of Aveiro, Portugal*
[2] *Agrarian School of Coimbra, Portugal*

Received 5 May 2013; Accepted 8 October 2013
Editor: Paulo Canas Rodrigues

Abstract The mapping defined by inter-nucleotide distances (InD) provides a reversible numerical representation of the primary structure of DNA. If nucleotides were independently placed along the genome, a finite mixture model of four geometric distributions could be fitted to the InD where the four marginal distributions would be the expected distributions of the four nucleotide types. We analyze a finite mixture model of geometric distributions (f_2), with marginals not explicitly addressed to the nucleotide types, as an approximation to the InD. We use BIC in the composite likelihood framework for choosing the number of components of the mixture and the EM algorithm for estimating the model parameters. Based on divergence profiles, an experimental study was carried out on the complete genomes of 45 species to evaluate f_2. Although the proposed model is not suited to the InD, our analysis shows that divergence profiles involving the empirical distribution of the InD are also exhibited by profiles involving f_2. It suggests that statistical regularities of the InD can be described by the model f_2. Some characteristics of the DNA sequences captured by the model f_2 are illustrated. In particular, clusterings of subgroups of eukaryotes (primates, mammalians, animals and plants) are detected.

Keywords genomic analysis; inter-nucleotide distances; geometric distributions; DNA

DOI: 10.19139/soic.v1i1.6

[*] Correspondence to: Department of Mathematics, University of Aveiro, Portugal. Email: adelaide@ua.pt

ISSN 2310-5070

1. Introduction

From the perspective of molecular evolution, DNA sequences can reflect both random mutation and selective evolution ([17]). In a simple way, DNA sequences are non-numerical sequences of the four-letter alphabet, A, C, G and T, which stands for the four nucleotides: Adenine, Cytosine, Guanine, and Thymine. Various transformations of DNA sequences into numerical data have been proposed in order to take advantage of methodologies available for quantitative data ([2, 9, 15, 16, 3, 18, 1]). Free-alignment algorithms have been applied to build distance trees aimed at visualizing historical evolutionary relationships among species (e.g., [21, 23]).

Numerical transformations that can capture useful information about mathematical properties discriminative and sensitive enough of variations in DNA composition and, at the same time, highlight important structural features of DNA sequences are desirable. Several numerical transformations of DNA sequences have been used to perform multiple organism comparisons. Basically, observed DNA sequences and randomly ordered sequences (random background) are compared using different procedures and discrepancy measures based either on genomic symbol frequencies (e.g., [22, 14, 13, 20]) or on genomic symbol distance frequencies (e.g., [1]). This type of residual analysis can highlight the contribution of DNA selective evolution of each species ([17]). In general, the random background behavior has a simple description in terms of probability distribution. For example, the random background for InD can be described using geometric distributions (e.g., [5, 1, 12]).

Usually, the random background of each species is translated by a independence model where each genomic symbol (e.g., nucleotide, dinucleotide) in the DNA sequences is assumed to be generated independently of the others, with occurrence probability estimated by its corresponding observed frequency. Using discrepancy-based methods, all studies mentioned above have shown that comparisons between the empirical distribution and the random background can be useful in detecting features (i.e., statistical patterns) of DNA sequences, as well as in obtaining (genomic) profiles for the differentiation of species. Nevertheless, discrepancy values have not been investigated from the estimation point of view. We extend the comparison-based approach such that divergence patterns can be estimated using a theoretical distribution.

In the present study, we describe three *approximation* models of the empirical distribution of the inter-nucleotide distances (InD) based on hypothetical evolutive arguments and independence assumption. The InD transforms any DNA sequence into a unique numerical sequence with the same length, such that each number represents the distance of a nucleotide to the next occurrence of the same nucleotide. If we assume that nucleotides are independently placed along a DNA sequence, then the InD can be fitted by a 4-component geometric mixture distribution for which the occurrence probability of each nucleotide type may be estimated by its corresponding observed frequency (random background). Furthermore, if the four nucleotides were identically distributed along the genome, the InD could be fitted by a geometric distribution with parameter $p = 1/4$ (baseline). Otherwise, and still under independence assumption, if a greater structural complexity of the InD in DNA sequences (e.g., different statistical features of the nucleotides associated to different regions of DNA) is intended for modelling, a g-component geometric mixture distribution could be suggested to reflect such complexity, with g to be determined.

Since the InD provides a reversible representation of DNA sequences, a probabilistic model fitted to the InD should reflect both the complexity, in terms of the species evolution, and the correlation structure of the nucleotides in the DNA sequences. The basic idea of taking mixture models of geometric distributions is to find a simple model suited to some structural complexity contained in InD data sets, under the independence assumption of the placement of nucleotides along the DNA sequences, and to investigate whether (divergence) patterns observed in the InD can also be established using such theoretical model.

In order to analyze the three approximations mentioned above, an experimental study is carried out using the DNA sequences of 45 different species. We analyze discrepancy values obtained when the estimated mixture, the random background, the baseline and the empirical distribution are compared for each species. Two different expressions for quantifying divergence between two distributions are considered separately. Our results show that, though the proposed mixture model does not fit to the InD, it may predict genome-wide characteristics based on discrepancy patterns. It is not surprising that independence-based models for InD do not fit all DNA structures, which are very complex systems.

The main contributions of this paper are: (i) the definition of a simple probabilistic model for each species, which contains information about statistical characteristics of the InD and allows the identification of biologically expectable clusterings of organisms and (ii) the exploitation of a divergence-based procedure to assess the fitting of probabilistic models to data sets.

The rest of this paper is organized as follows. In Section 2, besides formally defining the InD, we explain our motivation for investigating finite mixture models of geometric distributions for the InD and report a geometric mixture model to the InD according to the bayesian information criteria (BIC). The methodology used to investigate the ability of the proposed mixture model to capture information about the DNA sequences is described in Section 3. Section 4 presents the results of the experimental study involving the complete genomes of 45 species. Finally, in Section 5, we summarize the procedure used and the conclusions obtained.

2. Inter-nucleotide distance

2.1. Definition

Consider the alphabet $\mathcal{A} = \{A, C, G, T\}$ and let $V_i = \{1, 2, \ldots, i\}$ be the set of the first i positive integers. Taking the definition considered by [1], the InD is a mapping that transforms a nucleotide sequence of length L, $(S_1, S_2, \ldots, S_L)$ with $S_i \in \mathcal{A}$ for all i, into a numerical sequence $(d_1, d_2, \ldots, d_L)$ such that

$$d_i = \begin{cases} L - i + \min\{d \in V_i : \ S_d = S_i\}, \\ \qquad\qquad\qquad \text{if } \forall\, d \in V_{L-i} : \ S_{i+d} \neq S_i \\ \\ \min\{d \in V_{L-i} : \ S_{i+d} = S_i\}, \ \text{otherwise} \end{cases} . \tag{1}$$

For instance, for sequence *ATGATCGG*, the InD sequence is (3,3,4,5,5,8,1,3). Equation (1) assumes every linear DNA sequence as circular. Hence, for each nucleotide type, four artificial distances (from the last to the first nucleotide) are included in the mapping of each DNA sequence. It implies that the following properties are satisfied, in opposition to the definition of InD proposed by [15]:

Prop 1. $\sum_{i=1}^{L} d_i = \sum_{N\in\mathcal{A}} \sum_{i:\ S_i=N} d_i$ and, if the nucleotide N exists in the sequence $(S_1, S_2, \ldots, S_L)$, then $\sum_{i:S_i=N} d_i = L$;

Prop 2. Given the positions of the first occurrence of the four nucleotides in a DNA sequence, $s_1(N)$ with $N \in \mathcal{A}$, and the InD sequence $(d_1, d_2, \ldots, d_L)$, the corresponding nucleotide sequence $(S_1, S_2, \ldots, S_L)$ can be reconstructed iteratively. Indeed,

$$S_i = \arg\min_{N\in\mathcal{A}} s_i(N) \quad , \ i = 1, 2, \ldots, L \ ,$$

where $s_i(N)$ depends on S_{i-1} and is defined recursively by

$$s_i(N) = \begin{cases} d_{i-1} + s_{i-1}(N) & , \ N = S_{i-1} \\ s_{i-1}(N) & , \ N \neq S_{i-1} \end{cases} \tag{2}$$

for $i = 2, \ldots, L$.

2.2. Geometric finite mixture

Let D be the InD of a nucleotide in a certain position in the DNA sequence. Let P_N denote the (unknown) proportion of the nucleotide N in the DNA sequence ($\sum_{N\in\mathcal{A}} P_N = 1$), and let μ_N denote the mean distance between two consecutive N-type nucleotides, $N \in \mathcal{A}$. By the total probability law, D is modeled by a 4-component mixture distribution in the following manner

$$P(D = d) = \sum_{N\in\mathcal{A}} h_N(d)\, P_N \ , \quad d = 1, 2, \ldots \ , \tag{3}$$

where the marginal $h_N(d) \doteq P\,(D = d|N)$ is the probability distribution of the distance between two consecutive N-type nucleotides, $N \in \mathcal{A}$

If the distance between two consecutive N-type nucleotides was identically distributed for the four nucleotide types (say $h_N(.) = h(.)$, $\forall N \in \mathcal{A}$), then

$$P\,(D = d) = h(d)\ , \quad d = 1, 2, \dots\ .$$

If the nucleotide sequences were generated by serially independent random variables, then the distance between two consecutive N-type nucleotides could be modeled by a geometric distribution with parameter $p_N = 1/\mu_N$ and, consequently, D would follow a 4-component geometric mixture distribution defined by

$$P\,(D = d) = \sum_{N \in \mathcal{A}} p_N (1 - p_N)^{d-1}\, P_N\ , \quad d = 1, 2, \dots\ , \tag{4}$$

where the four marginal distributions are the corresponding expected distributions of the four nucleotide types.

In model (4) it is implicitly assumed that when looking for N-type only nucleotides along the DNA sequences, the success probability p_N is the same for every position of the genome. In case different success probabilities can be taken into account (this is biologically acceptable, for instance, when μ_N can vary for different regions of DNA sequences), the distances between two consecutive N-type nucleotides would be defined by a g_N-component geometric mixture model with probability distribution

$$h_N(d) = f(d; \Psi_N) \doteq \sum_{m=1}^{g_N} \pi_{N,m}\, p_{N,m} (1 - p_{N,m})^{d-1}\ , \tag{5}$$

with $d = 1, 2, \dots$ and parameter vector $\Psi_N \doteq (\pi_{N,1}, \pi_{N,2}, \dots, \pi_{N,g_N - 1}, p_{N,1}, p_{N,2}, \dots, p_{N,g_N})$, where (i) $0 \leq p_{N,m}, \pi_{N,m} \leq 1$, for $m = 1, \dots, g_N$, and $N \in \mathcal{A}$; (ii) $\pi_{N,m}$, $m = 1, \dots, g_N$, are related to the weights of the different success classes in the mixture such that $\sum_m \pi_{N,m} = 1$, for $N \in \mathcal{A}$; and (iii) the number g_N of mixture components reflects some structural complexity of the InD for the N-type nucleotides in the DNA sequences.

Thus, under the independence assumption of the nucleotides in the DNA sequences, from (3) and (5), the probability distribution of D could be defined by a mixture of four finite mixtures of geometric distributions involving a total number of $2G - 1$ one-dimensional parameters, where $G = \sum_{N \in \mathcal{A}} g_N$. Additionally, if the four nucleotide types are identically distributed (parameter vector Ψ_N is the same for all $N \in \mathcal{A}$, say $\Psi_N = \psi_g \doteq (\pi_1, \pi_2, \dots, \pi_{g-1}, p_1, p_2, \dots, p_g)$), a similar probability distribution for D, now with a lower total number of parameters

$(2g-1)$, would be provided:

$$\begin{aligned} P\left(D=d\right) &= \sum_{N\in\mathcal{A}} f(d;\psi_g)\, P_N = f(d;\psi_g) \\ &= \sum_{m=1}^{g} \pi_m\, p_m (1-p_m)^{d-1}\ , \quad d=1,2,\ldots . \end{aligned} \tag{6}$$

From a mathematical point of view, (6) could be interpreted as a general simple distribution probability which contains model (4) and model (3) satisfying (5). On the other hand, from a biological point of view, (6) can be interpreted as a general expected theoretical model of the InD under the assumption that all nucleotides are independently placed along the DNA sequences, such that its marginal distributions are now not explicitly addressed to the nucleotide types. Motivated by these two interpretations, in this work we analyze the approximation defined by (6) to the InD.

2.3. *Estimation*

We investigate the g-component geometric mixture model (6) as an approximation to the empirical distribution of the InD, where the parameter vector ψ_g will be estimated from the observed numerical sequences of InD resultant from a given genome. Remark that, after a value d in any numerical InD sequence, the value $d-1$ can not be found since if a nucleotide N is observed at position x and the next occurrence of the same nucleotide N is observed at position $x+d$, then the next occurrence of the nucleotide observed at position $x+1$ cannot be also found at position $x+d$ in the numerical sequence (for instance, in the sequence $NXXXNX$, where N represents one nucleotide and X represents a nucleotide that is different from N, the first InD is 4 but the second InD can never be 3 since $X \neq N$). In other words, the successive elements of the InD sequence obtained from the genetic sequences are not independent. Hence, the parameter vector ψ_g will be estimated using the composite likelihood framework ([10]). The methodology of composite likelihood reduces the dimension of the likelihood function using low-dimensional likelihood objects defined over subsets of data.

Let $\tilde{d} = (d_1, d_2, \ldots, d_L)$ denote the observed numerical sequence of InD from the DNA sequences of a given species. To form the composite likelihood, we rewrite $\tilde{d}$ taking into account the distances between the same nucleotide N-type, for $N \in \mathcal{A}$, as

$$\tilde{d} = \left(\tilde{d}^A, \tilde{d}^C, \tilde{d}^G, \tilde{d}^T\right)\ ,$$

where $\tilde{d}^N = \left(d_1^N, d_2^N, \ldots, d_{L_N}^N\right)$, $N \in \mathcal{A}$, denote the vector of observations sampled independently from the set of distances between N-type nucleotides and d_i^N denotes the ith observation contained in $\tilde{d}$ concerning with N-type nucleotides.

Thus, according to [10], the composite likelihood function will be defined as

$$\mathcal{CL}(\psi_g;\tilde{d}) = \prod_{N\in\mathcal{A}} \mathcal{L}_N(\psi_g;\tilde{d}) \ ,$$

where $\mathcal{L}_N(\psi_g;\tilde{d}) = f(\tilde{d}^N;\psi_g)$ is the marginal likelihood with respect to the distance between the same N-type nucleotides. Due to the first property of the InD mentioned above (Prop 1, in Section 2.1), we take the first $L_N - 1$ components of $\tilde{d}^N$ as identically and independently distributed observations and the last term $d^N_{L_N}$ is fixed. Thus,

$$\begin{aligned}
\mathcal{L}_N(\psi_g;\tilde{d}) &= f(d^N_1, d^N_2, \cdots, d^N_{L_N};\psi_g) \\
&= f(d^N_1, d^N_2, \cdots, d^N_{L_N-1}|d^N_{L_N};\psi_g) f(d^N_{L_N};\psi_g) \\
&= f(d^N_1, d^N_2, \cdots, d^N_{L_N-1};\psi_g) \times 1 \ , \text{ iff } \sum_{i=1}^{L_N} d^N_i = L \\
&= \prod_{i=1}^{L_N-1} f(d^N_i;\psi_g)
\end{aligned}$$

and, consequently, from (6),

$$\mathcal{CL}(\psi_g;\tilde{d}) \ = \ \prod_{N\in\mathcal{A}} \prod_{i=1}^{L_N-1} \sum_{m=1}^{g} \pi_m \, p_m (1-p_m)^{d^N_i - 1}$$

To find the maximum composite likelihood estimate $\widehat{\psi}_g$, we apply the EM algorithm ([6, 11]) using 10^{-5} as the tolerance value for the logarithm of the composite likelihood function $\mathcal{CL}(\psi_g;\tilde{d})$.

To determine g, we use the BIC approximation in the composite likelihood framework derived by [8] from the usual BIC method defined by [19]. BIC allows the selection of the number of mixture components needed to provide the best approximation to the density, penalizing the number of parameters in the model. Hence, g is the lowest positive integer that minimizes the function

$$BIC(g) = -2\log \mathcal{CL}(\widehat{\psi}_g;\tilde{d}) + \nu \log(L) \ ,$$

where $\nu = 2g - 1$ is the number of free parameters of the model (6).

3. Methodology

Several collections of sequenced genomes are currently available in public databases (e.g., Ensembl, NCBI). Usually, only one complete sequenced genome per sequenced species or sequenced strain is provided. DNA sequences available in public databases are assumed to be genome representatives of the sequenced species or strain ([7]). All the intrinsic properties of the DNA sequences associated to individuals belonging to a certain species are assumed to be contained in the DNA sequence, $(S_1, S_2, \ldots, S_L)$. Supposing that the corresponding InD sequence $(d_1, d_2, \ldots, d_L)$ represents a sample of D, statistical inference (confidence intervals and hypothesis testings) could be done to assess the fit of D by the model (6). We pay attention on three particular models by species:

1. $f_0(d) \equiv f(d; \psi_4)$ with $\pi_i = p_i = 1/4$, $i = 1, 2, 3, 4$. All the four nucleotides are supposed to be uniformly distributed in the DNA sequence. Under this model, DNA primary structure features of all organisms are assumed to be invariantly described by the same probability law; it will be denoted as the *baseline model*.
2. $f_1(d) \equiv f(d; \widehat{\psi}_4)$ with $\pi_i = p_i$, where p_i, $i = 1, 2, 3, 4$, are estimated by the relative frequencies of each of the four nucleotides in the DNA sequence. This model is similar to f_0 but takes into account the quantity of each nucleotide needed to describe each organism. It is the baseline specific of the DNA selective evolution of the species and represents the background random of the DNA sequence. Remark that f_1 coincides with (4) when p_N is replaced by its maximum likelihood estimate.
3. $f_2(d) \equiv f(d; \widehat{\psi}_g)$ with g determined by BIC as described in Section 2.3. It is expected that this model contains information about the structural complexity of the InD and so represents an approximation to the InD empirical distribution provided by an extension of the mixture model f_1.

Herein, we do not intend to conduct any statistical test. Our proposal is the construction of profiles based on divergences between the empirical distribution of the InD and each of the three models mentioned above. To measure the divergence between two probability functions $f_{k'}$ and f_k, at a point d, we use the score

$$s_{f_{k'}, f_k}(d) = \frac{f_{k'}(d) - f_k(d)}{\sqrt{\frac{f_k(d)(1 - f_k(d))}{L}}} \tag{7}$$

and the relative error

$$e_{f_{k'}, f_k}(d) = \frac{f_{k'}(d) - f_k(d)}{f_k(d)} \ . \tag{8}$$

The score (7) is motivated by the well-known test statistic used to test the null hypothesis $H_0 : P(D = d) = f_k(d)$, when $f_{k'}(d)$ is the observed relative frequency $f_{obs}(d)$ of the event $\{D = d\}$. The relative error (8) is motivated by [1]. In that paper, the authors compared the empirical distribution of the InD with the random background and created a genomic signature of a species, given by the

100-dimensional residue profile $[-e_{f_1,f_{obs}}(1), -e_{f_1,f_{obs}}(2), \cdots, -e_{f_1,f_{obs}}(100)]$. The choice of the vector dimension was arbitrary. When $f_{obs}(d) = 0$, for some d, those authors assigned the value zero to the relative error.

Unlike the score (7), the relative error (8) does not depend on the genome length L, and is related with (7) by the following equation:

$$s_{f_{k'},f_k}(d) = \frac{e_{f_{k'},f_k}(d)}{\sqrt{\frac{1}{L}\left(\frac{1}{f_k(d)} - 1\right)}} .$$

Both statistics (7) and (8) reflect a non-symmetrical measure of the divergence between f_k and $f_{k'}$ and can be used to construct profiles. Several symmetrical measures defined by commutative operations of (7) and (8) (e.g., $s_{f_{k'},f_k}(d) + s_{f_{k'},f_k}(d)$) can be easily defined. Independently of what is the formula that we fix, (7) or (8), the divergence measure value, at a point d, will be denoted by $r_{f_{k'},f_k}(d)$. Thus, for each species, we define a divergence profile given by the vector

$$\mathbf{r}_{f_{k'},f_k} = [r_{f_{k'},f_k}(1), r_{f_{k'},f_k}(2), \ldots, r_{f_{k'},f_k}(M)]$$

with $f_{k'}, f_k \in \{f_{obs}, f_0, f_1, f_2\}$, $k \neq k'$, where M is the lowest positive integer satisfying $f_{obs}(d) > 0, \forall d \in \{1, 2, \cdots, M\}$ and $f_{obs}(M + 1) = 0$.

Given the sequenced genome of n different organisms, we construct the divergence profile matrices

$$\mathbf{D}_{f_{k'},f_k} = [r_{f_{k'},f_k}(d; i)]_{d=1,2,\ldots,M; i=1,2,\ldots,n},$$

where $r_{f_{k'},f_k}(d; i)$ denotes the divergence measure value $r_{f_{k'},f_k}(d)$ for the species i. These matrices are useful for analyzing differences between any two distributions among n species. Fixing two generic distributions h_1 and h_2, a comparison between the divergence profile matrices $\mathbf{D}_{h_1,f_k}$ and $\mathbf{D}_{h_2,f_k}$ allows the identification of similar divergence patterns for the two distributions h_1 and h_2 among the n species. An analogous reasoning can be made using $\mathbf{D}_{f_k,h_1}$ and $\mathbf{D}_{f_k,h_2}$. Obviously, since non-symmetrical measures of divergence are here considered, the divergence profile matrices $\mathbf{D}_{f_k,h_1}$ and $\mathbf{D}_{h_2,f_k}$ are not comparable. When the divergence profiles between the empirical distribution (f_{obs}) and the baseline model (f_0), or the random background (f_1), are considered, relations among species in terms of evolutionary pattern (i.e., clustering with close phylogenetic relationships) can be highlighted. Graphically, both traditional hierarchical clustering techniques and principal component analysis can provide useful tools for visualizing those divergence similarities among the n species. By using the former, dendrograms can be constructed that resemble a kind of distance trees. By using the latter, projections of the divergence profiles on a two-dimensional space defined by the first two principal components can be displayed, providing another different manner to visually explore divergence

Table 1. Notation of the ten euclidean distance matrices constructed from matrices $\mathbf{D}_{g_1,g_2}$ for different choices $g_1, g_2 \in \{f_{obs}, f_0, f_1, f_2\}$, $g_1 \neq g_2$. Since divergences between models f_0 and f_1 are not of interest to this study, the euclidean distance matrices $(+)$ associated to $\mathbf{D}_{f_0,f_1}$ and $\mathbf{D}_{f_1,f_0}$ are not considered. Note that there is no symmetric behavior since $r_{f_{k'},f_k}(d) \neq r_{f_k,f_{k'}}(d)$.

	g_2			
g_1	f_{obs}	f_0	f_1	f_2
f_{obs}		c_0	c_1	c_2
f_0	a_0		+	A_0
f_1	a_1	+		A_1
f_2	a_2	C_0	C_1	

patterns involving f_{obs} and the theoretical models f_0, f_1 and f_2. If f_2 is an approximation *acceptable* to f_{obs}, then the marginal distributions can be associated to (homogeneous) classes underlying to the empirical distribution of the InD and the number of the mixture components can be useful as an initial number on methods for identifying clusters ([4]).

In order to evaluate the approximation f_2 to f_{obs} we propose assessing the similarity of divergence patterns drawn by two comparable divergence profiles involving the distributions f_{obs}, f_2, f_1 and f_0. In order to measure the similarities, we suggest obtaining the matrix $(n \times n)$ of the euclidean distances between each pair of rows (species) for each divergence profile matrix, transform each one into a n^2–dimensional vector and calculate the Kendall coefficient of concordance between the two n^2-dimensional vectors obtained. The Kendall coefficient value is an estimate of the overall agreement between the two euclidean distance matrices, regardless of the exact values of the euclidean distance between divergence profiles. An higher level of concordance implies an higher similarity between the two euclidean distance matrices, and so an higher similarity between the two correspondent dendrograms constructed using any linkage criteria based on order statistics (e.g., single and complete linkages). Following this divergence-based approach, the levels of concordance between the ten euclidean distance matrices obtained from matrices $\mathbf{D}_{f_{obs},f_k}$, $\mathbf{D}_{f_k,f_{obs}}$, for $k = 0, 1, 2$, and $\mathbf{D}_{f_2,f_k}$, $\mathbf{D}_{f_k,f_2}$, for $k = 0, 1$, sketched in Table 1, are aimed at quantifying divergence between the empirical distribution f_{obs} and the approximation f_2.

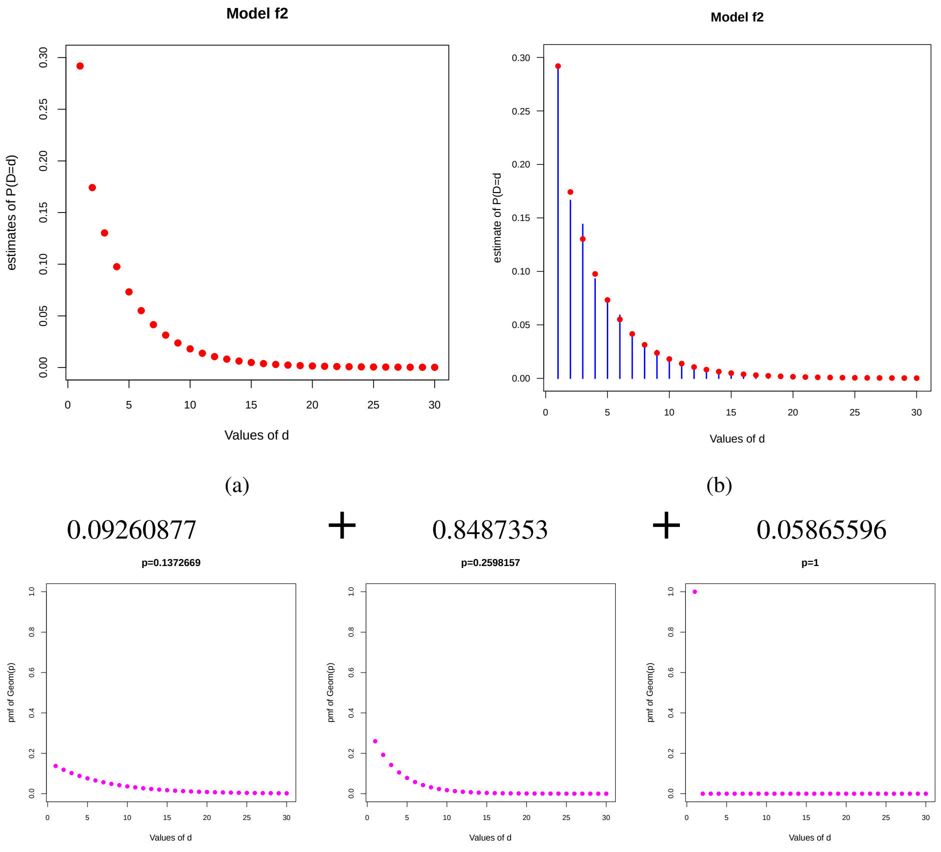
4. Results

Based on the methodology described above, we evaluated whether the proposed mixture model f_2 contains information about f_{obs} and thus reveals important features of the DNA sequences. To perform this evaluation, the assembled DNA sequences of a set of 45 species were collected from several public databases: 42 from the National Center for Biotechnology Information (NCBI, ftp://ftp.ncbi.nih.gov/genomes/); 1 from the Joint Genome Institute (http://genome.jgi-psf.org/), 1 from Xenbase (http://www.xenbase.org/); and other from Genome Project (http://www.fugu-sg.org/). All symbols in each of the collected sequences that did not correspond to one of the four nucleotides were removed before further processing. The names of the species considered are listed in Appendix (Table 6).

The species under analysis are 5 archaea, 13 bacteria and 27 eukaryotes. Among the eukaryotes, 4 protozoa, 3 fungi, 4 plants and 16 animals are considered. Analyzing the genome size of each species, differentiation among groups is highlighted: prokaryotes and fungi present a shorter genome size than that of protozoa, which, in turn, is shorter than that of plants and animals (data not shown).

Taking into account the observed InD sequences, the finite mixture models f_0, f_1 and f_2 were estimated for each species, as mentioned in Section 3 (the estimated parameter vector Ψ_g is not shown). Figure 1 illustrates the model f_2 estimated for the species *Streptococcus pneumoniae* (St). Figure 2 depicts the choice made for the number g of components associated to the model f_2 and based on BIC (see Table 5, in Appendix, for more detailed results). While organisms with a shorter genome size are associated to lower complexity models ($g \leq 4$), organisms with a larger genome size (e.g., mammals) are associated to higher complexity models ($g \geq 4$, except for one species: *Oryza sativa*). Concretely, the estimated mixture models with a lower number of components arise more tendentiously associated to the prokaryotes than to the eukaryotes, suggesting that the genome of prokaryotes may be characterized by a lower number of homogeneous classes than that of eukaryotes. Also, a flat shape of the BIC is observed in Figure 2, for all species. This means that an higher number of marginal geometric distributions could be incorporated in the model f_2 without greatly affecting the balance established by BIC between the likelihood model and its complexity. From a biological perspective, this fact suggests that there may be additional classes governing the InD sequences in genomes, but the identification of those potentially new classes is probably outside the reach of our model f_2.

A preliminary analysis of the maximum likelihood estimates $\widehat{\Psi}_g$ obtained for the 45 species was carried out. A summary description of all the estimates $\widehat{p}_m$, $m = 1, \cdots, g$, is presented in Table 2. The estimates $\widehat{p}_m$ associated to the geometric distribution with the highest mixture weight in the mixture model f_2 are highlighted in Figure 3. Two interesting features of the model f_2 were observed. The first one is that the estimates for the parameter of the geometric distribution with the highest mixture weight are close to the value 0.25 for several species (Figure 3). These values may indicate that a large weight of the model



(a) (b)

(c) (d) (e)

Figure 1. Mixture model f_2 estimated for the species *Streptococcus pneumoniae* (St). On the top, red points represent the estimated model f_2 and blue bars represent the empirical model f_{obs}. On the bottom, the $g = 3$ geometric distributions and their corresponding weights in the mixture model f_2 are depicted. Hence, for the species St, the model f_2 is defined by a mixture with 9% of a geometric distribution with parameter $p_1 = 0.137$, 85% by a geometric distribution with a parameter close to 0.25 ($p_2 = 0.260$) and an extra weight of 6% for the calculation of $P(D = 1)$.

f_2 is associated to randomness of the four nucleotides in the DNA sequence and, consequently, the ability of model f_2 to explain some important features of the DNA sequences can be questioned. This fact was already expected (e.g., the correlation structure of the data was not explicitly considered in the mixture model (6)). The second interesting feature is that for various species, the estimates

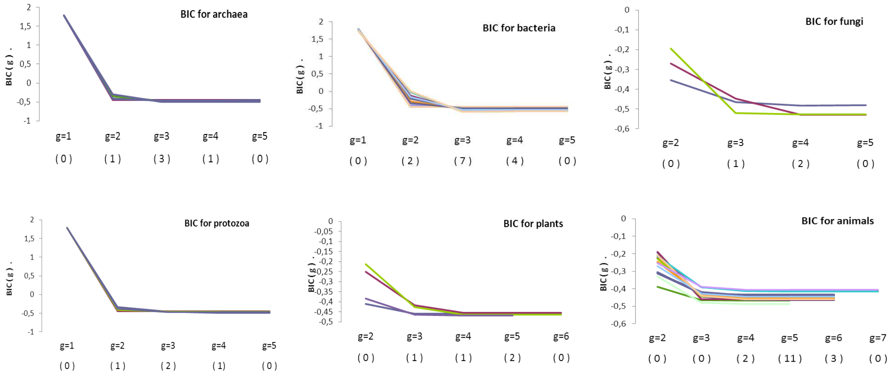


Figure 2. Graphical representation of the normalized BIC values obtained for each species. Each line is associated to one species. The total number of species for which each g is the lowest minimizer of the function $BIG(g)$, and consequently the total number of species approximated by a g-component mixture model f_2, is shown in parenthesis.

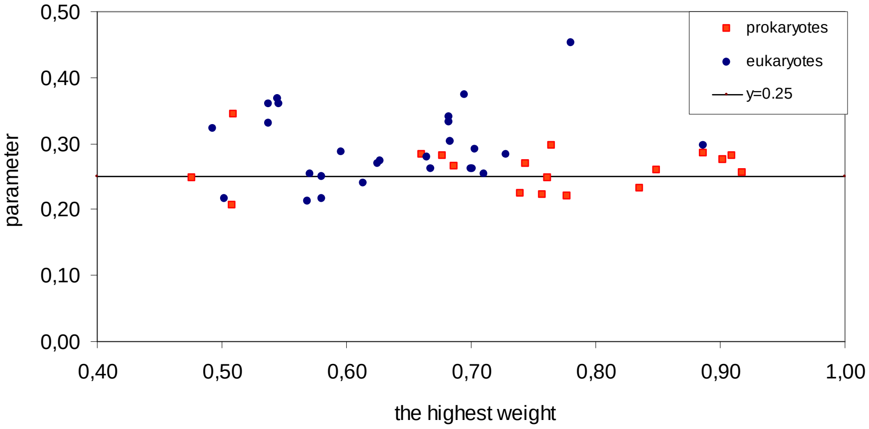


Figure 3. Parameter value of the geometric distribution with the highest weight $(\max \pi_i)$ observed in the proposed mixture model f_2 estimated for each species.

yielded by the EM algorithm led to $\widehat{p}_m = 1$, for some m (Table 3). This fact highlights the importance of the existence of pairs of equal nucleotides in the DNA sequences provided by model f_2 for several species.

Among the 45 species under study, the lowest observed value of the InD with null frequency was $d = 55$ (exhibited by the species *Chlamydia trachomatis*). Hence, following the methodology described in Section 3, for each divergence

Table 2. Mean and standard deviation of the parameters of the geometric distributions sorted by the weight in the mixture model f_2. Here $\widehat{p}_{(i)}$ represents the estimate of parameter of the geometric distribution related with the i-th lower weight in the mixture model f_2 for each species.

		$g = 2$	$g = 3$	$g = 4$	$g = 5$	g=6
Prokaryotes	$\widehat{p}_{(1)}$	0.13±0.01	0.56±0.47	0.08±0.03		
	$\widehat{p}_{(2)}$	0.28±0.005	0.38±0.27	0.64±0.49		
	$\widehat{p}_{(3)}$		0.25±0.03	0.37±0.22		
	$\widehat{p}_{(4)}$			0.27±0.05		
Eukaryotes	$\widehat{p}_{(1)}$	0.11	0.06±0.02	0.03±0.03	0.01±0.01	0.01±0.00
	$\widehat{p}_{(2)}$	0.30		0.32±0.35	0.06±0.03	0.07±0.02
	$\widehat{p}_{(3)}$			0.17±0.03	0.37±0.36	0.47±0.29
	$\widehat{p}_{(4)}$			0.32±0.06	0.38±0.16	0.30±0.26
	$\widehat{p}_{(5)}$				0.25±0.03	0.28±0.01
	$\widehat{p}_{(6)}$					0.28±0.01

Table 3. Mean and standard deviation of the estimated weights $\widehat{\pi}_m$ associated to the estimates $\widehat{p}_m = 1$ obtained for the model f_2 by organism groups in study. N is the total number of species where was observed $\widehat{p}_m = 1$, for some m, in the estimated model f_2.

	group	N	mean	standard deviation
Prokaryotes	bacteria	8	0.06	0.03
Prokaryotes	archaea	0	–	–
Eukaryotes	fungi	3	0.05	0.01
Eukaryotes	protozoa	1	0.12	–
Eukaryotes	plants	2	0.04	0.004
Eukaryotes	animals	1	0.04	–

formula (7) and (8), we drew the ten euclidean distance matrices depicted in Table 1 with $M = 54$ and $n = 45$.

By applying hierarchical clustering techniques with complete and single linkage and euclidean distance as similarity index to various divergence profile matrices, $\mathbf{D}_{f_{k'},f_k}$, and divergence profile matrices based on symmetrical statistics (e.g., $s_{f_{k'},f_k}(d) + s_{f_{k'},f_k}(d)$), several dendrograms as distance trees of the 45 species were obtained. In all of these dendrograms, some scientifically accepted evolutionary relationships between living organisms were partially detected. For instance, the separation of prokaryotes, fungi and protozoa from plants and animals. Also, a mammalian cluster and a primate cluster were detected in accordance with the scientifically accepted phylogeny (e.g., [20]).

Principal component analysis for $2n \times 54$ augmented divergence profile matrices

$$[D_{f_{obs},f_k} \| D_{f_2,f_k}]' , \quad k = 0, 1 ,$$

with $n = 45$ (all the species), $n = 18$ (restricted to the 18 prokaryotes) and $n = 27$ (restricted to the 27 eukaryotes) aimed at facilitating graphical comparisons between divergence profiles $\mathbf{r}_{f_{obs},f_k}$ and $\mathbf{r}_{f_2,f_k}$, for $k=0,1$, through their projections on the same reduced two-dimensional space of the two first principal components, were performed. Let's focus on these projections. In Figure 4 the projections for the prokaryotes and the eukaryotes are separately displayed. Plots very similar to the graphs (c) and (d) in Figure 4, with all the prokaryotes projected over the fungi group, were obtained when the 45 species was displayed (graphs not shown). Most projections of $\mathbf{r}_{f_2,f_k}$ are not overlapped with the corresponding projections of $\mathbf{r}_{f_{obs},f_k}$, $k=0,1$, particularly in the case of the prokaryotes. Although this fact shows that f_{obs} and f_2 are different, they share several interesting features: i) in terms of divergence with f_1, the ratio of the difference between the first principal component of the projection of $\mathbf{r}_{f_{obs},f_1}$ and that of $\mathbf{r}_{f_2,f_1}$ to the difference between the second principal component of the projection of $\mathbf{r}_{f_{obs},f_1}$ and that of $\mathbf{r}_{f_2,f_1}$ seems very similar for all prokaryotic organisms (parallel lines joining red and blue points in Figure 4 (b)); ii) in terms of divergence with f_k, $k=0,1$, the first principal component of the projections of $\mathbf{r}_{f_{obs},f_k}$ and that of $\mathbf{r}_{f_2,f_k}$ are visually very close to one another for each eukaryotic organism (vertical lines joining red and blue points in Figure 4 (c),(d)); it is also true for prokaryotic organisms only when the divergence is measured in terms of the model f_0 (Figure 4 (a)); iii) in terms of clusterings, from the projections of $\mathbf{r}_{f_2,f_k}$ and $\mathbf{r}_{f_{obs},f_k}$, $k=0,1$, several clusters of eukaryotes are clearly detected in accordance with the scientifically accepted evolutionary processes: the fungi cluster (*Sp, Sc, Ca*), the primate cluster (*Mu, Pt, Hs*), and the mammalian cluster (*Bt, Eq, Pt, Mm, Hs, Cf, Oa, Rn, Mu*), being this latter linearly separable from the other species by an oblique (vertical, resp.) line when divergences of both f_{obs} and f_2 are measured from f_0 (f_1, resp.). These similarities between divergence patterns involving f_2 and f_{obs} show that the model f_2 contains information about the structural features of the InD in the DNA sequence, being sufficiently robust to enable the identification of clusters of organisms, particularly eukaryotic groups. In Figure 5 the clusterings mentioned above are highlighted.

For all species, we detected high divergence values $r_{f_{obs},f_k}(d)$ and $r_{f_k,f_{obs}}(d)$, $k=0,1,2$, for several values of d. Therefore, the InD is not fitted to the models f_k, $k=0,1,2$. To assess the unfitting of the empirical distribution f_{obs} to f_2, we applied the procedure described in Section 3 with the notation introduced in Table 1. Concretely, we calculated the Kendall coefficient of concordance between the euclidean distance matrices a_0, a_1 and a_2 (c_0, c_1 and c_2, resp.) associated to $\mathbf{D}_{f_k,f_{obs}}$ ($\mathbf{D}_{f_{obs},f_k}$, resp.) and the euclidean distance matrices A_0 and A_1 (C_0 and C_1, resp.) associated to $\mathbf{D}_{f_k,f_2}$ ($\mathbf{D}_{f_2,f_k}$, resp.). Table 4 summarizes some of the results obtained. The levels of concordance were calculated considering not only the whole set of $n=45$ species under study but also subsets of $n=10,20,30$ species randomly sampled without replacement from those 45 species. Independently of the value n and the divergence formula considered, (7) and (8), the results clearly show higher levels of concordance between euclidean distance

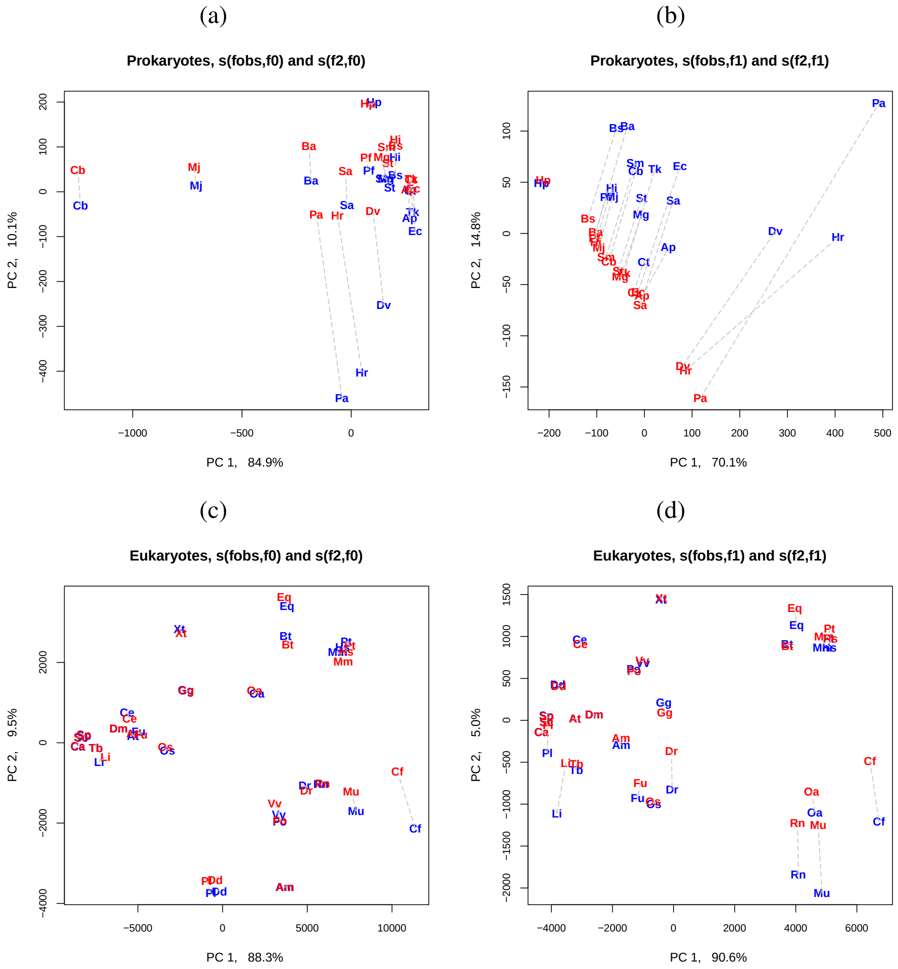


Figure 4. Plots of the first and second principal components for the $2n \times 54$ augmented divergence profile matrices $\left[D_{f_{obs},f_k} \| D_{f_2,f_k}\right]'$ ((a) and (c) for $k = 0$, (b) and (d) for $k = 1$) restricted to 18 prokaryotes -(a), (b)-, and restricted to 27 eukaryotes -(c), (d)-. In the four graphs, the projections of the divergence profiles $\mathbf{r}_{f_{obs},f_k}$ and $\mathbf{r}_{f_2,f_k}$ are depicted in blue and red, respectively. The percentage of variability explained by each component is indicated in the axis labels.

matrices denoted by capital letters with euclidean distance matrices denoted by the corresponding lowercase letters: a_0 and A_0, a_1 and A_1, c_0 and C_0, and c_1 and C_1. For different pairs above, the highest levels of concordance are close to 100% ($\geq 94.9\%$ for the score s and $\geq 82.5\%$ for the relative error e). These highly

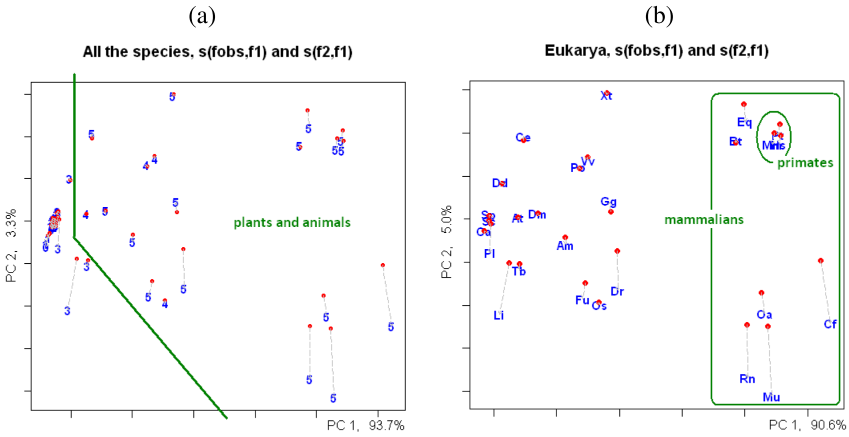


Figure 5. Similiar plots as depicted in Figure 4 for the $2n \times 54$ augmented divergence profile matrices $\left[D_{f_{obs},f_1} \| D_{f_2,f_1}\right]'$ for all species -(a)- and restricted to 27 eukaryotes -(b)-. Here the detected clusterings are highlighted.

Table 4. Kendall coefficient values ($\times 100$) to quantify the level of concordance between the euclidean distance matrices indicated in Table 1. In parentheses are the mean and the standard deviation of ten Kendall coefficient values obtained when the procedure is repeated ten times with 10 species sampling without replacement from the set of 45 species.

	residue s		residue e	
	A_0	A_1	A_0	A_1
a_0	95.6	86.1	82.5	30.1
	(94.1±3.8)	(85.1±5.8)	(79.3±7.9)	(31.4±15.2)
a_1	83.6	94.9	23.4	86.8
	(80.6±8.2)	(94.9±4.5)	(20.6±15.9)	(85.7±6.2)
a_2	52.4	56.3	35.5	18.2
	(52.3±10.1)	(57.2±8.1)	(21.1±19.4)	(14.5±22.7)
	C_0	C_1	C_0	C_1
c_0	97.1	70.5	89.2	-2.5
	(97.1±1.7)	(69.3±9.2)	(90.3±4.9)	(-4.2±13.7)
c_1	69.7	95.7	-4.2	90.5
	(67.4±10.1)	(95.4±4.6)	(-7.1±11.4)	(91.2±4.5)
c_2	51.5	59.1	-1.8	-11.1
	(52.2±13.0)	(60.2±7.3)	(-5.3±10.2)	(-11.2±10.2)

concordant euclidean distance matrix pairs correspond to the divergence profile matrices $\mathbf{D}_{f_k,f_{obs}}$ and $\mathbf{D}_{f_k,f_2}$, and to divergence profile matrices $\mathbf{D}_{f_{obs},f_k}$ and $\mathbf{D}_{f_2,f_k}$, $k = 0, 1$. These results show that possible relations among the species in terms of divergence profiles from random and baseline models can still be captured by the model f_2.

5. Conclusions

There are different ways of investigating properties of DNA sequences. Herein, we considered the InD, a reversible numerical representation of DNA sequences, and explored its fitting by finite geometric mixture models. Based on the DNA sequences of 45 species, we analyzed divergence between the InD empirical distribution and the three mixture models: (i) f_0, which represents a probable starting model where all organisms are invariantly described by this same probability law; (ii) f_1, which represents the background random model of the DNA sequence in its selective evolution; and (iii) f_2, which represents a simple likelihood model suited to some structural complexity contained in the InD, under independence assumption of the nucleotides in the DNA sequences, where the number of parameters was selected in accordance with a composite likelihood version of the BIC and the parameters were estimated by the EM algorithm.

Using two formulas to quantify the divergence between two probability functions $f_{k'}$ and f_k, (7) and (8), we defined different divergence profile matrices for $n = 45$ species. Each profile depends on a function pair $f_{k'}, f_k \in \{f_{obs}, f_0, f_1, f_2\}$, with $k \neq k'$.

The observation of the divergence profile matrix $\mathbf{D}_{f_{obs},f_2}$ allowed us to check the unfitting of the empirical distribution f_{obs} to the model f_2.

Although the proposed mixture model f_2 is still not adequate for modeling the InD, our experimental analysis, based on divergence patterns from the models f_0 and f_1, indicts the existence of various common features between the empirical distribution of InD and the model f_2, namely, i) similar relations between projections of divergence profiles on a reduced two-dimensional space, and ii) similar clusterings (Figure 4 and Table 4). For the 45 species considered and from a biological point of view, the model f2 is able

- to predict a discrimination of the eukaryotes organisms for mammalian and non-mammalian groups when divergence profiles of these organisms are represented in a two-dimensional reduced space (Figure 5(b)).
- to predict the separation of plants and animals from other species when divergence profiles of all the species are represented in a two-dimensional reduced space (see Figure 5(a))
- to highlight an additional weight for the occurrence of dinucleotides with equal nucleotides in the DNA sequences, for several species (for an example, see Figure 1).

All these facts lead us to conclude that the model f_2 is able to capture information from DNA sequences.

In addition, regarding the mixture model f_2 estimated for each species, its number of mixture components and the identification of its corresponding marginal distributions may provide useful insights on biological mechanisms, unveiling the existence and the probabilistic structure of homogeneous classes underlying the DNA sequences. In particular, there is a tendency for the number of mixture components of the model f_2 to be greater in eukaryotes than in prokaryotes.

Using the InD, we explored the idea of comparing the proposed mixture model f_2 with the empirical distribution f_{obs} after removing the random background (f_1) and the baseline (f_0) from both of these distributions, in order to evaluate the ability of the model f_2 for capturing information of DNA sequences. A similar procedure could be extended to other theoretical models and mappings of DNA sequences.

Acknowledgement

This work was supported by *FEDER* funds through *COMPETE*–Operational Programme Factors of Competitiveness ("Programa Operacional Factores de Competitividade") and by Portuguese funds through the CIDMA, *Center for Research and Development in Mathematics and Applications* (University of Aveiro), and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER-022690.

A. Supplementary information

	$g = 2$	$g = 3$	$g = 4$	$g = 5$	$g = 6$
Prokaryotes	Dv, Hr, Pa	Ap, Ct, Ec, Hp, Mg, Pf Sa, Sm, St, Tk	Ba, Bs, Cb, Hi, Mj		
Eukaryotes	Li	Pl, Tb, Os	Ca, Sc, Sp Dd, At Eq, Fu	Po, Vv, Am, Bt, Ce, Dm, Dr, Gg, Mm, Mu, Pt, Rn, Xt	Cf, Hs, Oa

Table 5. Identification of the number of components g of the mixture model f_2 determined by BIC for each of the 45 species in study.

Species	Abbreviation	Reference
Aeropyrum pernix	*Ap*	NC000854
Halobacterium salinarum	*Hr*	NC010364, NC010366, NC010369
Methanococcus jannaschii	*Mj*	NC000909, NC001732, NC001732
Pyrococcus furiosus	*Pf*	NC003413
Thermococcus kodakarensis	*Tk*	AP006878
Bacillus anthracis	*Ba*	NC003997
Bacillus subtilis	*Bs*	NC000964
Chlamydia trachomatis	*Ct*	NC000117
Clostridium botulinum	*Cb*	NC009495, NC009496
Desulfovibrio vulgaris	*Dv*	NC008741, NC008751
Escherichia coli	*Ec*	NC000913
Haemophilus influenzae	*Hi*	NC000907
Helicobacter pylori	*Hp*	NC000915
Mycoplasma genitalium	*Mg*	NC000908
Pseudomonas aeruginosa	*Pa*	NC002516
Staphylococcus aureus	*Sa*	NC002951, NC006629
Streptococcus mutans	*Sm*	NC004350
Streptococcus pneumoniae	*St*	NC011900
Arabidopsis thaliana	*At*	AGI 7.2
Oryza sativa	*Os*	NC008394, NC008405
Populus trichocarpa	*Po*	Build 1.0
Vitis vinifera	*Vv*	Build 1.1
Bos taurus	*Bt*	Build 4.1
Cannis familiaris	*Cf*	Build 2.1
Equus caballus	*Eq*	Build 2.1
Gallus gallus	*Gg*	Build 2.1
Apis mellifera	*Am*	Build 4.1
Drosophila melanogaster	*Dm*	Build 4.1
M musculus	*Mu*	Build 37.1
Caenorhabditis elegans	*Ce*	NC003279
Rattus norvegicus	*Rn*	Build 4.1
Xenopus Tropicalis	*Xt*	Build 4.1
Homo sapiens	*Hs*	Build 36.3
Macaca mulatta	*Mm*	Build 1.1
Pan troglodytes	*Pt*	Build 2.1
Danio rerio	*Dr*	Build 3.1
Takifugu rubripes	*Fu*	fourth assembly
Ornithorhynchus anatinus	*Oa*	Build 1.1
Dictyostelium discoideum	*Dd*	Build 2.1
Leishmania infantum	*Li*	NC009277, NC009386, NC009420
Plasmodium falciparum	*Pl*	Build 2.1
Trypanosoma brucei	*Tb*	NC005063, NC007276, NC007283, NC007334, NC008409, NT165287:88
Candida albicans	*Ca*	NC007436
Saccharomyces cerevisiae	*Sc*	SGD 1
Schizosaccharomyces pombe	*Sp*	Build 1.1

Table 6. List of the species considered in the present study, their abbreviations and the DNA builds used.

REFERENCES

1. Vera Afreixo, Carlos A. C. Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo J. S. G. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23):3064–3070, December 2009.

2. Vera M. A. Afreixo, Paulo J. S. G. Ferreira, and Dorabella M. S. Santos. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, 14(6):523–530, November 2004.
3. Mahmood Akhtar and Julien Epps. Signal processing in sequence analysis: Advances in eukaryotic gene prediction. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):310–321, June 2008.
4. J-P. Baudry, A. Raftery, G. Celeux, K. Lo, and R. Gottardo. Combining mixture components for clustering. *Journal of Computational and Graphical Statistics*, 19(2):332?53, 2010.
5. P Carpena, P Bernaola-Galván, R Román-Roldán, and J.L Oliver. A simple and species-independent coding measure. *Gene*, 300(1-2):97 – 104, 2002.
6. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
7. E. S. Lander, et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.
8. X. Gao and P. X.-K. Song. Composite likelihood bayesian information criteria for model selection in high dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540, 2010.
9. Bo Liao. A 2d graphical representation of DNA sequence. *Chemical Physics Letters*, 401:196–199, 2005.
10. B. Lindsay. Composite likelihood methods. In *Statistical inference from stochastic processes*, pages 221–239, Providence, RI: American Mathematical Society, 1988.
11. Geoffrey McLachlan and Krishnan. *The EM algorithm and extensions*. John Wiley, 2008.
12. Katharina Mir, Klaus Neuhaus, Siegfried Scherer, Martin Bossert, and Steffen Schober. Predicting statistical properties of open reading frames in bacterial genomes. *PLoS ONE*, 7(9):e45103, 09 2012.
13. G. Moura, M. Pinheiro, J. Arrais, A.C. Gomes, L. Carreto, A. Freitas, J.L. Oliveira, and M. Santos. Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mrna primary structure. *PLoS ONE*, 2(9):e847, July 2007.
14. G. Moura, M. Pinheiro, Raquel M. Silva, Isabel M. Miranda, Vera M. A. Afreixo, GD Gaspar Dias, Adelaide Freitas, J. L. Oliveira, and Manuel Santos. Comparative context analysis of codon pairs on an ORFeome scale. *Genome Biology*, 6(3):R28(14), 2005.
15. A. S. S. Nair and T. Mahalakshmi. Visualization of genomic data using inter-nucleotide distance signals. In *Proceedings of IEEE Genomic Signal Processing*, 2005.
16. Miguel Pinheiro, Vera Afreixo, Gabriela Moura, Adelaide Freitas, Manuel A. Santos, and J.L. Oliveira. Statistical, computational and visualization methodologies to unveil gene primary structure features. *Methods of Information in Medicine*, 45:163–168, 2006.
17. Ji Qi, Bin Wang, and Bai-Iin Hao. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *Journal of molecular evolution*, 58:1–11, 2004.
18. Milan Randic and Jure Zupan. On representation of DNA by line distance matrix. *Journal of Mathematical Chemistry*, 43(2):674–692, 2008.
19. Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.
20. Gregory E Sims, Se-Ran Jun, Guohong Albert Wu, and Sung-Hou Kim. Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8):2677–2682, February 2009.
21. Gregory E Sims, Se-Ran Jun, Guohong Albert Wu, and Sung-Hou Kim. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences of the United States of America*, 106(40):17077–17082, October 2009.
22. F. Takeuchi, Y. Futamuraa, H. Yoshikurac, and K. Yamamoto. Statistics of trinucleotides in coding sequences and evolution. *Journal of Theoretical Biology*, 222:139–149, 2003.
23. S. Vinga and J. Almeida. Alignment-free sequence comparison – a review. *Bioinformatics*, 19(4):513–523, 2003.