Exploring Homology Using the Concept of Three-State Entropy Vector

Armando J. Pinho[1], Sara P. Garcia[1], Paulo J. S. G. Ferreira[1], Vera Afreixo[2], Carlos A. C. Bastos[1], António J. R. Neves[1], and João M. O. S. Rodrigues[1]

[1] Signal Processing Lab, DETI / IEETA,
University of Aveiro, 3810–193 Aveiro, Portugal
[2] Department of Mathematics,
University of Aveiro, 3810–193 Aveiro, Portugal
{ap,spgarcia,pjf,vera,cbastos,an,jmr}@ua.pt

Abstract. The three-base periodicity usually found in exons has been used for several purposes, as for example the prediction of potential genes. In this paper, we use a data model, previously proposed for encoding protein-coding regions of DNA sequences, to build signatures capable of supporting the construction of meaningful dendograms. The model relies on the three-base periodicity and provides an estimate of the entropy associated with each of the three bases of the codons. We observe that the three entropy values vary among themselves and also from species to species. Moreover, we provide evidence that this makes it possible to associate a three-state entropy vector with each species, and we show that similar species are characterized by similar three-state entropy vectors.

Keywords: DNA signature, DNA coding regions, DNA entropy, Markov models

1 Introduction

It is well-known that there are periodicities in DNA sequences, the strongest of which is generally associated with the period three that can be found in the exons of prokaryotes and eukaryotes [**?**,**?**]. This three-base periodicity has been used, for example, for predicting potential protein-coding regions [**?**,**?**,**?**,**?**,**?**] and for finding potential reading frame shifts in genes [**?**].

In a previous work [**?**,**?**], we have used this property for exploring the possibility of using a three-state finite-context model with the aim of improving the compression of the protein-coding regions of the DNA sequences. That study led us to the conclusion that, for those protein-coding DNA regions, a model that switches sequentially between three states provides better compression than a model based on a single state. Moreover, the three-state model looses its efficacy when applied to unrestricted DNA sequences, which provides additional evidence towards the distinctive three-base periodicity of the protein-coding regions.

Besides the observation that a three-state finite-context model works better than a single-state model in protein-coding regions, we also observed a phenomenon that caught our attention. Each of the three states of the finite-context

model can be viewed as a model of the information source associated to each of the three nucleotides that form a codon. Since we are able to estimate the entropy of each of the three states of our model, we are also able to estimate the average information carried out by each of the three nucleotides. The interesting finding that we have made was that both the absolute and the relative values of these entropies vary among the species [**?**,**?**]. In other words, the average information conveyed when the first, second or third bases of a codon are specified is not the same, and the differences depend on the species.

In this paper, we further investigate this phenomenon and, particularly, we try to find out if the differences among the values of the entropies of the three base positions of the codon could be used as a species signature. Although still preliminary, the results obtained suggest that this is in fact true, i.e., that we are able to construct a low-dimensional entropy vector capable of correctly clustering similar species. Therefore, these findings may contribute to the development of new methods for alignment-free sequence comparison.

2 Materials and Methods

2.1 DNA Sequences

In this preliminary study, we used thirteen species, nine eukaryotes (five animals and four plants) and four prokaryotes (bacteria), listed in Table 1. When available, we used the RNA data provided in a single file. In the other cases, we used the data of the ".ffn" files. In the case of the *Ricinus communis* we used the ".cds" data. Because the performance of the three-state model is affected by losses of synchronization in the reading frame, i.e., it assumes that, for example, the first base of the codon is always handled by state one of the model, we only considered sequences whose length is a multiple of three and that do not contain undefined symbols. Moreover, for these experiments, and also with the aim of avoiding inconsistencies in the expected codon structure, we did not consider those that do not start with ATG.

2.2 Finite-Context Models

Consider an information source that generates symbols, s, from the alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Consider that the information source has already generated the sequence of n symbols $x^n = x_1 x_2 \ldots x_n,\ x_i \in \mathcal{A}$. A finite-context model (see Fig. 1) assigns probability estimates to the symbols of the alphabet, regarding the next outcome of the information source, according to a conditioning context computed over a finite and fixed number, $k > 0$, of the most recent past outcomes $c = x_{n-k+1} \ldots x_{n-1} x_n$ (order-k finite-context model) [**?**,**?**,**?**]. Therefore, the number of conditioning states of the model is 4^k.

The probability estimates, $P(X_{n+1} = s|c), \forall_{s \in \mathcal{A}}$, are usually calculated using symbol counts that are accumulated while the sequence is processed, which makes them dependent not only of the past k symbols, but also of n. In other

Table 1. Organisms used in this study.

Organism	Reference
Homo sapiens (human)	Build 37.1
Pan troglodytes (chimpanzee)	Build 2.1
Macaca mulatta (rhesus macaque)	Build 1.1
Mus musculus (mouse)	Build 37.1
Rattus norvegicus (brown rat)	Build 4.1
Arabidopsis thaliana (thale cress)	NC003070/1/4/5/6
Populus trichocarpa (black cottonwood)	Version 2.0
Vitis vinifera (grape vine)	Build 1.1
Ricinus communis (castor oil plant)	Release 0.1
Streptococcus pneumoniae strain ATCC 700669	NC011900
Chlamydia trachomatis strain D/UW-3/CX	NC000117
Mycoplasma genitalium strain G-37	NC000908
Streptococcus mutans strain UA159	NC004350

words, these probability estimates will in general vary as a function of the position along the sequence.

Typically, the probability estimates produced by the finite-context model are used to drive an arithmetic encoder, which is able to generate output bit-streams with average bitrates almost identical to the entropy of the model [**?**,**?**,**?**]. The theoretical bitrate average of the finite-context model after encoding n symbols is given by

$$H_n = -\frac{1}{n}\sum_{i=1}^{n} \log_2 P(X_i = x_i|c) \quad \text{bpb}, \tag{1}$$

where $c = x_{i-k} \ldots x_{i-2}x_{i-1}$ and "bpb" stands for "bits per base". Recall that the entropy of any sequence of four symbols is limited to two bits per symbol, a value that is obtained when the symbols are independent and equally likely.

The probability that the next outcome, X_{n+1}, is s, where $s \in \mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, is obtained using the estimator

$$P(X_{n+1} = s|c) = \frac{n_s^c + \alpha}{n^c + \alpha|\mathcal{A}|}, \tag{2}$$

where n_s^c represents the number of times that, in the past, the information source generated symbol s having as conditioning context $c = x_{n-k+1} \ldots x_{n-1}x_n$ and where

$$n^c = \sum_{s \in \mathcal{A}} n_s^c \tag{3}$$

is the total number of events that has occurred so far in association with context c. The parameter α controls how much probability is assigned to possible but yet unseen events. In this work, we used $\alpha = 1$, which transforms the estimator into the multinomial extension of Laplace's rule of succession [**?**].

Note that, initially, when all counters are zero, the symbols have probability 1/4, i.e., they are assumed equally probable. The counters are updated each time

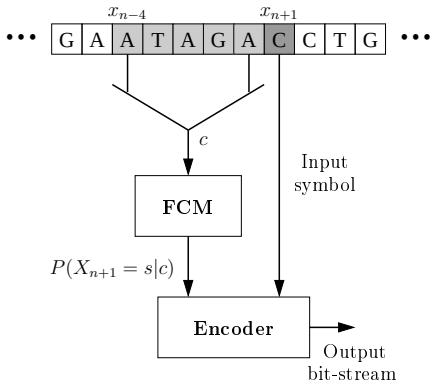

Fig. 1. Example of a finite-context model: the probability of the next outcome, X_{n+1}, is conditioned by the k last outcomes. In this example, $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and $k = 5$. The "Encoder" block is usually an arithmetic encoder.

a symbol is encoded. Since the context template is causal, the decoder is able to reproduce the same probability estimates without needing additional information. In other words, this model is self-contained, in the sense that it is capable of recovering the original sequence based only on the bit-stream produced by the encoder.

2.3 The Three-State Model

Figure 2 shows the model addressed in this paper. It differs from the finite-context model displayed in Fig. 1 by the inclusion of three internal states. Each state is selected periodically, according to a three-base period, and comprises a finite-context model, similar to the one presented in Fig. 1.

The three-state model, originally introduced in [**?**,**?**] with the purpose of compressing protein-coding regions of DNA, is revisited in this paper with the aim of exploring homology using the idea of a three-state entropy vector.

With this model, probabilities depend not only on the k last outcomes, but also on the value of (n mod 3), which is used for state selectivity. In this case, the probability estimator is given by

$$P(X_{n+1} = s|c) = \frac{n_s^{c,\phi} + \alpha}{n^{c,\phi} + \alpha|\mathcal{A}|}, \tag{4}$$

where

$$\phi = n \bmod 3 + 1 \quad \text{and} \quad n^{c,\phi} = \sum_{s \in \mathcal{A}} n_s^{c,\phi}. \tag{5}$$

Therefore, three different sets of counters are used, one for each state. Moreover, only the counters associated with the chosen state are updated. It is worth noting that, in order to be able to operate, this model does not require the

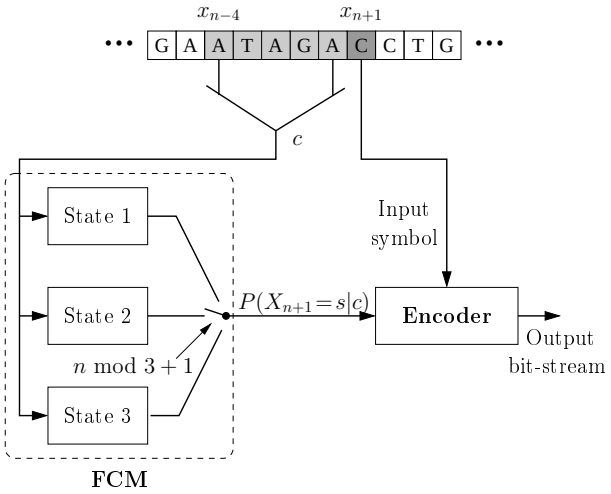

Fig. 2. Three-state model, exploiting the three-base periodicity of the DNA protein-coding regions. In this case, the probability of the next outcome, X_{n+1}, is conditioned both by the k last outcomes and by the value of ($n \bmod 3 + 1$).

knowledge of the correct reading frame. However, once a particular initial position has been chosen, the corresponding reading frame should be maintained, otherwise the statistics will become mixed and the model will not work properly. Notwithstanding, if we intend to determine the entropies associated with each of the three base positions inside the codons, we need to know which base position corresponds to each state of the model. Moreover, note that (1) needs to be modified accordingly, leading to the entropies

$$H_n^1 = -\frac{1}{\lfloor n/3 \rfloor} \sum_{i=1}^{\lfloor n/3 \rfloor} \log_2 P(X_{3i-2} = x_{3i-2}|c), \tag{6}$$

where $c = x_{3i-k-2} \ldots x_{3i-4} x_{3i-3}$,

$$H_n^2 = -\frac{1}{\lfloor n/3 \rfloor} \sum_{i=1}^{\lfloor n/3 \rfloor} \log_2 P(X_{3i-1} = x_{3i-1}|c), \tag{7}$$

where $c = x_{3i-k-1} \ldots x_{3i-3} x_{3i-2}$, and

$$H_n^3 = -\frac{1}{\lfloor n/3 \rfloor} \sum_{i=1}^{\lfloor n/3 \rfloor} \log_2 P(X_{3i} = x_{3i}|c), \tag{8}$$

where $c = x_{3i-k} \ldots x_{3i-2} x_{3i-1}$.

For the cases reported in this paper we always started the model at the beginning of a codon, implying that state one corresponds to the first base position of the codon, state two to the second base position and state three to the third base position.

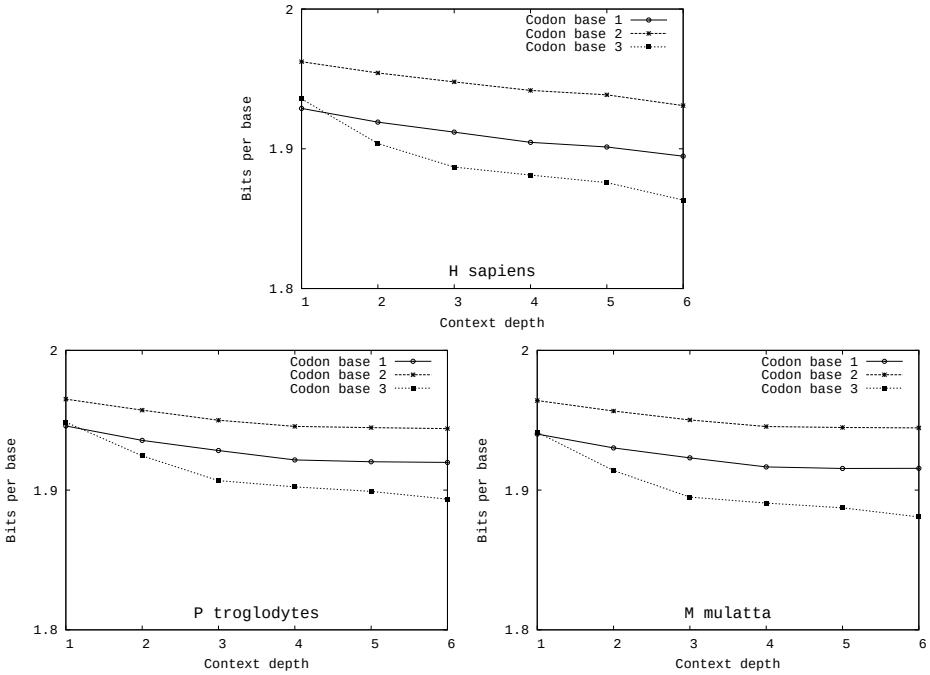

Fig. 3. Plots showing the distribution of the information among the three bases of the codon for *H. sapiens*, *P. troglodytes* and *M. mulatta*.

3 Results

We ran the three-state finite-context model for the DNA sequences under analysis, using contexts of depths one to six, i.e., from $k = 1$ until $k = 6$. Figures 3–6 display graphics of the average number of bits per base obtained. Each graph contains three curves, one for each of the three bases of the codon, i.e., the values of H_n^1, H_n^2 and H_n^3 after having processed the whole sequence.

As can be seen, the plots shown in Fig. 3, corresponding to the *H. sapiens*, *P. troglodytes* and *M. mulatta* organisms, present a significant similarity. Moreover, for most of the values of k (the depth of the context) the entropy associated to the second base of the codon is the largest, followed by the first and third bases.

This behavior is also observed in the graphs of Fig. 4, where the *M. musculus* and *R. norvegicus* organisms are addressed. However, in this case, and in contrast to the previous one, it can be seen a clear inversion of the entropy of the first and third bases for $k = 1$.

Figure 5 displays the entropy graphs for the four plants used in this preliminary assessment, namely the *A. thaliana*, *P. trichocarpa*, *V. vinifera* and *R. communis*. For these organisms, the entropy of the first base is generally larger than that of the second one, which is larger than the entropy of the third base.

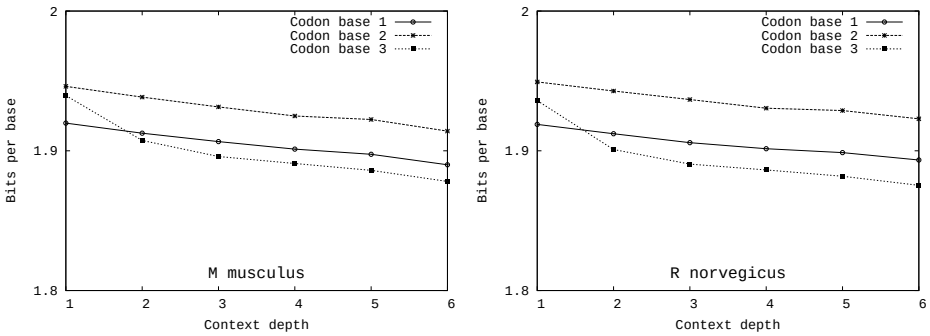

Fig. 4. Plots showing the distribution of the information among the three bases of the codon for the *M. musculus* and *R. norvegicus*.

Therefore, in comparison to the five animals, there is a change in the relative position of the curves regarding the first and second bases of the codon.

This same ordering can be found in the curves corresponding to the *S. pneumoniae*, *C. trachomatis*, *M. genitalium* and *S. mutans* organisms, presented in Fig. 6. However, whereas for the plants the difference between the values of the upper and lower curves is typically less than 0.05 bpb, in the case of the four bacteria this difference is typically larger than 0.1 bpb.

In order to better understand the similarities and differences of the entropy values among the analyzed species, we have built a dendogram (Fig. 7) with the PHYLIP package (`http://evolution.genetics.washington.edu/phylip.html`), constructed using the unweighted pair group method with arithmetic average (UPGMA), also known as average linkage method [**?**]. The distance matrix was obtained by computing the Euclidean distance between vectors built from the three-state entropy vectors corresponding to context depths from one to six. Therefore, each organism is represented by a vector with eighteen elements, i.e., the concatenation of six groups of three-state entropies.

Regarding this dendogram, we have some remarks. The prokaryotes (lower branch) are correctly separated from the eukaryotes (upper branch), except for the bacterium *C. trachomatis*. Amongst the prokaryotic branch, all bacteria are correctly grouped. The clades for the animals and plants are also well identified. Amongst the plants, *P. trichocarpa* should be classified closer to *R. communis*, as they belong to the same order. As for the animals, the human should be closer to the chimpanzee, then to the Rhesus macaque, and finally to the mouse and brown rat. Tough these minor misclassifications, this methodology correctly identifies overall clades, making these preliminary results encouraging in the exploration of three-state finite-context models for a meaningful classification of organisms.

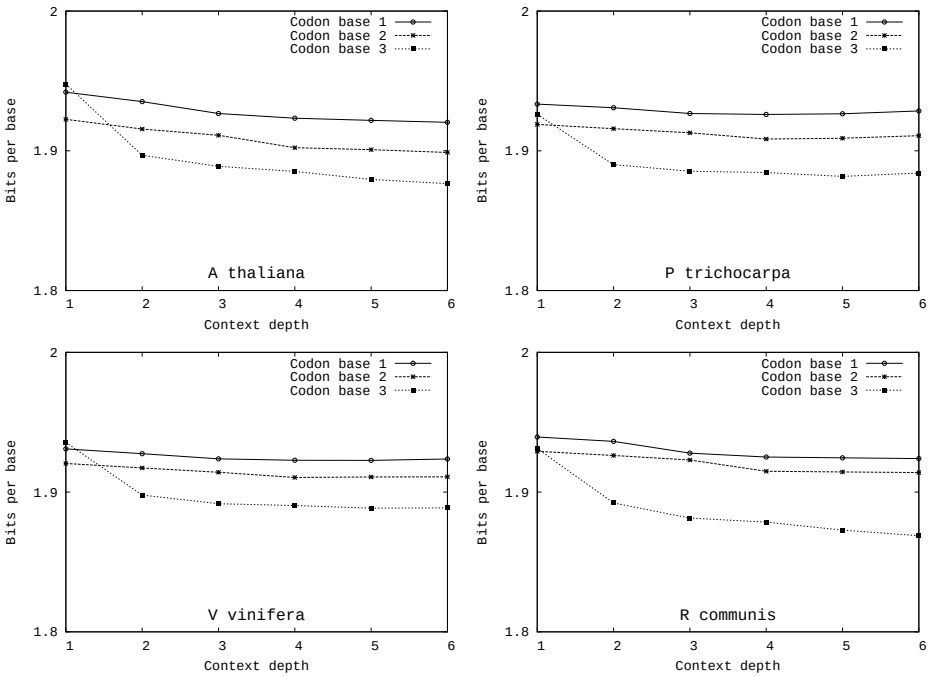

Fig. 5. Plots showing the distribution of the information among the three bases of the codon for *A. thaliana*, *P. trichocarpa*, *V. vinifera* and *R. communis*.

4 Conclusion

The three-base periodicity of the exons has been used since its discovery mostly as an aid in gene finding. More recently, it was noted that the three entropy values associated to each of the three base positions of the codon are not the same, and that the differences vary from organism to organism. We refer to these three entropy values as the "three-state entropy vector" of the organism.

The work presented in this paper is a start towards a deeper investigation of the implications of this observation, particularly in what concerns its use for species classification. The preliminary results obtained suggest that the information gathered from the three-state entropy vector alone seems to be sufficient for building meaningful dendograms, encouraging further study.

5 Acknowledgments

This work was supported in part by the grant with the COMPETE reference FCOMP-01-0124-FEDER-007252 (FCT, Fundação para a Ciência e Tecnologia, reference PTDC/EIA/72569/2006). Sara P. Garcia acknowledges funding from the European Social Fund and the Portuguese Ministry of Science, Technology and Higher Education.

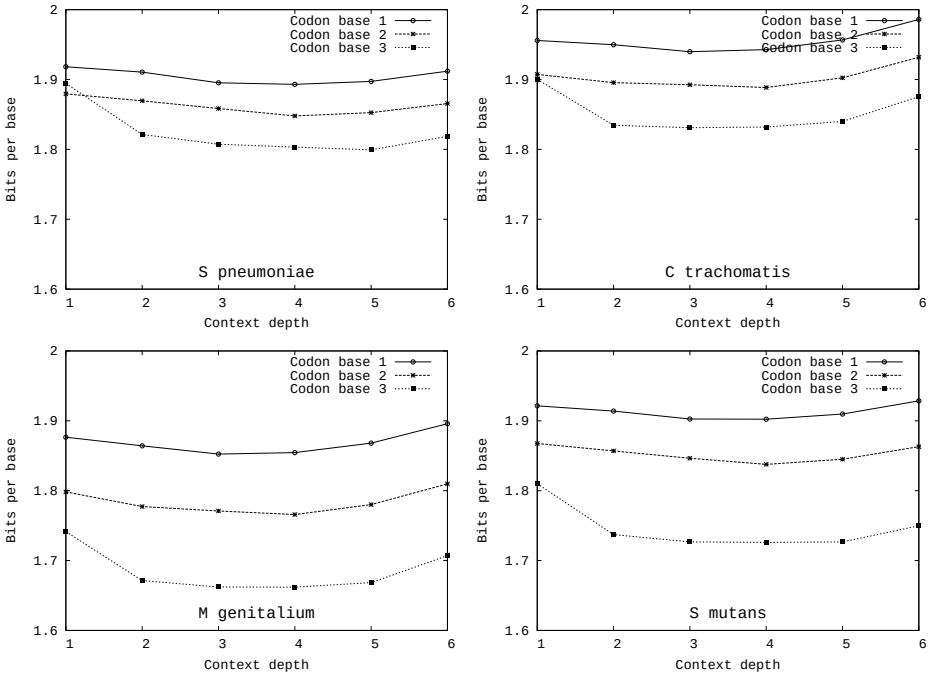

Fig. 6. Plots showing the distribution of the information among the three bases of the codon for *S. pneumoniae*, *C. trachomatis*, *M. genitalium* and *S. mutans*.

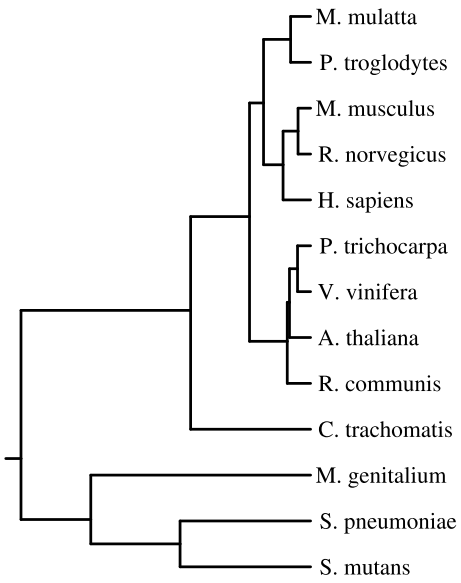

Fig. 7. Dendogram, based on UPGMA, obtained from the matrix of the Euclidean distance between the three-state entropy vectors for context depths from one to six.