Analysis of DNA sequences using finite-context modelling and compression*

Diogo Pratas and Armando J. Pinho
Signal Processing Lab, DETI / IEETA
University of Aveiro, 3810–193 Aveiro, Portugal
pratas@ua.pt, ap@ua.pt

Abstract

Models that rely exclusively on the Markov property, usually known as finite-context models, can model DNA sequences without considering mechanisms that take advantage of exact and approximate repeats. These models provide probability estimates that depend on the recent past of the sequence and have been used for data compression. In this paper, we investigate how finite-context modelling can be used to characterise the information sources associated to the DNA sequences. The experimental results, obtained using the human genome, show patterns generated by these models that can reveal important information about the DNA.

1. Introduction

The human genome is determined by approximately 3 000 million bases (A, C, G, T) [5]. This means that it takes approximately 750 MBytes to represent the human genome using $\log_2 4 = 2$ bits per symbol (bps).

Frequently, the main motivation for studying data compression algorithms is the need for efficient storage, as the state-of-the-art XM [1], or transmission of information. However, there is another aspect of paramount interest that is associated to compression algorithms: every compression method has an underlying data model that might unveil important characteristics of the data. Therefore, looking for DNA compression algorithms is also a form of finding models that describe the information source associated to DNA.

In this paper, we address a powerful modelling technique that has been used for compressing DNA sequences [3, 2, 4], the finite-context models, with the objective of finding patterns that might be a characteristic of certain DNA sequences.

*This work was supported in part by the grant with the COMPETE reference FCOMP-01-0124-FEDER-010099 (FCT, Fundação para a Ciência e Tecnologia, reference PTDC/EIA-EIA/103099/2008).

2. Results

The DNA sequences used for obtaining the results presented in this paper are from the release of April 14th, 2003 of the Human genome (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/April_14_2003). We have used a DNA compressor based on finite-context models that is described in [4].

For reference, we generated three independent and equally distributed pseudo-random sequences, using uniform probabilities (i.e., 0.25 probability for each of the four DNA bases), and with different sizes. We compressed these sequences using finite-context models with 16 different orders (context depths). The final entropy values have been plotted in Fig. 1.

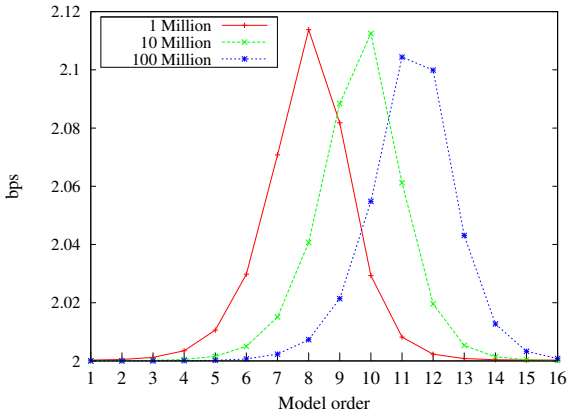

Figure 1. The entropy of synthetic DNA sequences with 1, 10 and 100 million bases, obtained using finite-context models of several orders (depths).

According to Fig. 1, it is possible to identify a property of the finite-context models which consists on the fact that the compression of random sequences produces a peak in the average number of bits per symbol (entropy) for an or-

der that depends on the size of the sequence. Moreover, more randomness in the sequence seems to be related with a better peak representation, a property that we use in this study.

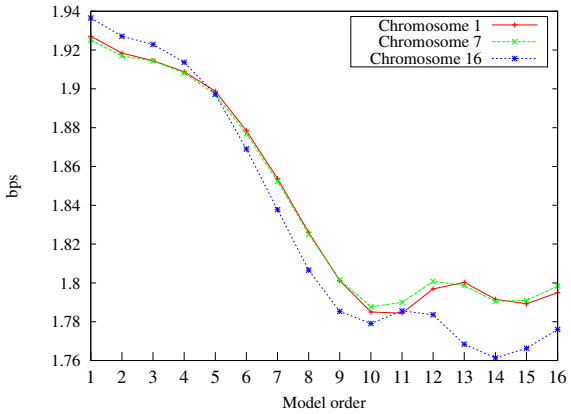

Figure 2. Entropy curve for human chromosomes 1, 7 and 16.

For comparing the behaviour of the entropy curves that we have obtained for random sequences with those generated for real DNA sequences, we have ran the same procedure for all human chromosomes. We have observed that almost all chromosomes have an identical pattern, although significantly different from that obtained with random sequences. This pattern can be observed for three examples of chromosomes in Fig. 2. In this case, the curves show a peak, although much less evident than it is for the random sequences. Based on this observation, we can conclude that there are parts in these chromosomes that seem to be random. Furthermore, the horizontal position of the peak shows that chromosome 1 (having about 247 million bases) is bigger than chromosome 7 (about 158 million bases), and that chromosome 7 is bigger than chromosome 16 (90 million bases), as we expected.

However, chromosomes 19, X and Y do not show an entropy pattern identical to the others, specially chromosome Y. As can be seen in Fig. 3, chromosomes 19 and X are generally better compressed than the others (excluding chromosome Y), revealing that there are more repetitive zones and less random parts in these chromosomes.

Chromosome Y (with about 60 million bases) showed the most different behaviour in this process. There is no peak in the entropy curve, as can be seen in Fig. 3. As we know from previous studies, the Y chromosome is highly repetitive, a property that agrees with the observations and that strengthens the conclusion that the main reason for peak absence is the existence of extensive repetitive zones in this chromosome.

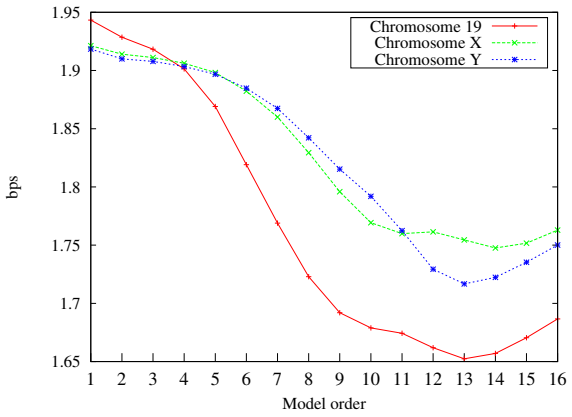

Figure 3. Entropy curve for human chromosomes X, 19 and Y.

3. Conclusions

Finite-context modelling has been used for DNA data compression. In this paper, we applied these models for DNA data analysis, using the information provided by the entropy curves obtained using different orders of the models. We verified that these models generate a peak in the entropy curve when the data is random. Based on this property, we studied the chromosomes of the human genome, concluding that the corresponding entropy curves differ significantly from the random ones and also that they differ among them according to the amount of repetitiveness of the sequence.

References

[1] M. D. Cao, T. I. Dix, L. Allison, and C. Mears. A simple statistical algorithm for biological sequence compression. In *Proc. of the Data Compression Conf., DCC-2007*, pages 43–52, Snowbird, Utah, Mar. 2007.

[2] A. J. Pinho, A. J. R. Neves, C. A. C. Bastos, and P. J. S. G. Ferreira. DNA coding using finite-context models and arithmetic coding. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP-2009*, pages 1693–1696, Taipei, Taiwan, Apr. 2009.

[3] A. J. Pinho, A. J. R. Neves, and P. J. S. G. Ferreira. Inverted-repeats-aware finite-context models for DNA coding. In *Proc. of the 16th European Signal Processing Conf., EUSIPCO-2008*, Lausanne, Switzerland, Aug. 2008.

[4] A. J. Pinho, A. J. R. Neves, D. A. Martins, C. A. C. Bastos, and P. J. S. G. Ferreira. Finite-context models for DNA coding. In S. Miron, editor, *Signal Processing*, pages 117–130. INTECH, Mar. 2010.

[5] L. Rowen, G. Mahairas, and L. Hood. Sequencing the human genome. *Science*, 278:605–607, Oct. 1997.