PhD Work Plan

Analysis of DNA sequences through compression-based complexity profiles

PhD student: Diogo Pratas (pratas@ua.pt)
Supervisor: Armando J. Pinho (ap@ua.pt)

Research Unit: Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA)

1 Abstract

The complexity profile of a DNA sequence is a numerical sequence of the same length that indicates a measure of the predictability of each DNA base. Complexity profiles allow, for example, looking for repetitive structures inside a chromosome or across several chromosomes, which are often associated with regulatory functions. The complexity profiles can also be used in finding evolutionary distances and, therefore, to build phylogenetic trees.

We will study the full potential of using compression-based complexity profiles for the analysis of long DNA sequences. This will be done along two lines: (1) intra-species analysis, where the goal is to explore the information conveyed by the complexity profiles for locating and classifying repetitive structures occurring inside a chromosome or across several chromosomes of the same species; (2) inter-species analysis, where our goal is to use the complexity information for computing evolutionary distances among species. Finite-context modeling will play a key role.

2 State-of-the-art

About fifteen years ago, Grumbach and Tahi [1] proposed the first dedicated DNA compression algorithm, Biocompress. Surely more important than the compression provided by Biocompress, which was rather modest, was the observation made by the authors that the development of DNA data compression algorithms is important not only for data reduction, but also, and in some cases even more important, in understanding the laws that govern the data.

This was not a novel observation, even in the context of DNA sequence processing and analysis. Other researchers had pointed out previously such important aspect, although sometimes implicitly (see, for example, [2–4]). The reason why we mention here explicitly the first DNA compression method is because DNA compression plays a central role in this project proposal, as we will explain shortly.

Several authors continued to point out the relevance of DNA data compression for DNA data analysis, such as pattern discovery and classification (see [5–7], for some examples). In fact, the motivation for using data compression and, consequently, the shortest possible description of the data, dates back at least to the works of Solomonoff, Kolmogorov, Chaitin and Wallace et al. in the mid 60's, and is generally known as the Kolmogorov complexity or algorithmic entropy.

Since the work of Grumbach and Tahi [1], several contributions have been made in the area of DNA data compression (see [8–11], for some of the most recent). In this context, the work of Allison et al. has been of particular interest, because they have been trying to relate the information content of a DNA sequence, by means of the per symbol code length generated by the encoder, with important characteristics of DNA, such as repetition structures. These information sequences, as they call them, were first presented in [12] and, more recently, in [7], where they propose using them for the comparative analysis of long DNA sequences.

We have recently shown that combinations of finite-context models are able to attain significant performance in DNA sequence compression [13–15]. We introduced new updating mechanisms, allowing these models to capture information regarding the inverted repeats usually found in DNA sequences and

we investigated several aspects related to multiple finite-context models that compete or cooperate for encoding the data [14, 15]. In conclusion, we have shown that DNA can be well represented by Markovian models.

One of the key advantages of DNA compression based on finite-context models is that the encoders are fast and have O(n) time complexity. In fact, most of the effort spent by previous DNA compressors is in the task of finding exact or approximate repeats of sub-sequences or of their inverted complements. No doubt, this approach has proved to give good returns in terms of compression gains, but normally at the cost of long compression times. Although slow encoders could be tolerated for storage purposes (compression could be ran in batch mode), for interactive applications, such as those that we will address, they are certainly not appropriate.

3 Objectives

The analysis of the huge amounts of genomic data that are continuously been generated puts a number of challenging problems to several research areas and, particularly, to the areas of computational biology and bioinformatics. One of these challenges is related with the problem of generating complexity profiles of DNA sequences in an efficient way, because these sequences might be as large as entire genomes (for example, the human genome is composed of about 3000 million bases).

The advantages of modeling DNA as combinations of finite-context models is twofold. On one hand, it shows that, contrarily to what was generally assumed, DNA can be well represented by Markovian models. On the other hand, and because they have O(n) time complexity, they can be applied to very long sequences in a time efficient way, a characteristic not present in the other competing DNA compression techniques.

We will use compression-based complexity profiles generated by finite-context models for studying biologically significant characteristics of DNA sequences, both within the chromosomes of a given species (intra-species analysis) and across chromosomes of different species (inter-species analysis). The intra-species analysis might be highly relevant for characterizing, for example, repetition structures that are known to be common in regulatory regions, both inside a given chromosome or across several chromosomes of the same species. Inter-species analysis aims at contributing for establishing distances among different species and, therefore, to build phylogenetic trees.

By the end of this work, we will have contributed with the development of new tools for DNA data analysis, fundamentally in aspects related to the discovery and characterization of functionally significant regions of DNA and in the classification of species.

4 Description

The construction and analysis of DNA complexity profiles has been an important topic of research, due to its applicability in the study of regulatory functions of DNA, comparative analysis of organisms, genomic evolution and others [16, 17]. For example, it has been observed that low complexity regions of DNA are often associated with important regulatory functions [18].

Several measures have been proposed for evaluating the complexity of DNA sequences, such as the Lempel-Ziv complexity, the linguistic complexity or compression-based complexity measures (see, for example, [19]). Among those, we find the compression-based approaches the most promising and natural, because compression efficiency is clearly defined (it can be measured by the number of bits generated by the encoder). In fact, the bitstream generated by compressing a given string over a given alphabet can be regarded as a program that, together with the appropriate decoder, can reproduce the original string. Hence, the size of that bitstream can be seen as an upper bound on the Kolmogorov complexity of the string. Therefore, the better the compression technique is, the tighter the bound will be.

In our ongoing work on DNA data compression, we have obtained several encouraging and important results related to DNA modeling based on finite-context models (also known as Markov models). Proba-

bly the most important of those results is the finding that DNA sequences can be much better represented by Markov models than what it was previously believed [13–15]. The finite-context models have been already able to produce preliminary complexity profiles that are almost identical to those produced by much more time consuming algorithms. In fact, the difference in time requirements is overwhelming. For example, compressing the human chromosome number 2 with the techniques based on finite-context models takes less than ten minutes in a 1.66 GHz laptop computer. This DNA sequence has about 240 million bases.

In this PhD project, we will study the full potential of using compression-based complexity profiles for the analysis of long DNA sequences. This will be done in two major, although related, research topics. One is intra-species analysis, where the major goal is to explore the information conveyed by the complexity profiles for locating and classifying repetitive structures occurring inside a chromosome or across several chromosomes of the same species. The complexity profiles can be generated for one or more chromosomes of a given species and be used for intra-species analysis, for example, for characterizing repetitive structures. Therefore, one of our aims is to automatically locate these low complexity regions and report them.

In Figure 1 of the Annex, we show some preliminary and exploratory entropy-based complexity profiles, obtained using finite-context models. In this example, we show the complexity profile of chromosome 1 of the Cyanidioschyzon merolae organism, a primitive red alga which lives in acidic hot water. In that graphic, we can see several regions where the complexity value goes well below the baseline level that, for an entropy-based complexity profile of DNA, can be set at two bits per DNA nucleotide. The two regions which we have marked with letters A and B correspond to telomeric inverted repeate sequences.

The other topic is inter-species analysis, where the main goal is to use the complexity information for computing evolutionary distances among the species. We expect that compressing DNA sequences with multiple finite-context models can be used to compute evolutionary distances, and with them building phylogenetic trees from inferred homologies [20]. The idea is to consider a given sequence, such as the exons of a protein-coding gene in the mouse genome, for training. Such sequence is coded with the best finite-context model for each sequential window of a given length (for instance, 100 bases), producing an information sequence with the ordered optimal finite-context models for coding that given sequence. Then, a chromosome of a different organism, e.g. human, is scanned for coding with such information sequence. From the overall complexity profile, a plateau of much lower entropy values will match the exact position of the human gene homologous to the mouse gene from which the information sequence was generated. Evolutionary distances will be inferred by quantifying the effect of the found homology on the complexity profile, i.e. by how much the entropy values decrease.

Figure 2 of the Annex illustrates the idea, using some preliminary and exploratory work that we have performed. The figure shows an entropy-based complexity profile obtained by scanning over a 1Mb window the human (Homo sapiens) chromosome 7 with finite-context models trained with the TAX1BP1 gene of the mouse (Mus musculus). For comparison, we drew (in green) the position of the exons of the homologous gene in the human chromosome, which, as can be seen, match the low complexity peaks of the profile. This type of homology can also be searched in non-protein-coding regions of the genome.

References

[1] S. Grumbach and F. Tahi, "Compression of DNA sequences," in *Proc. of the Data Compression Conf., DCC-93*, Snowbird, Utah, 1993, pp. 340–350.

[2] L. Allison and C. N. Yee, "Minimum message length encoding and the comparison of macromolecules," *Bulletin of Mathematical Biology*, vol. 52, pp. 431–431, May 1990.

[3] P. Salamon and A. K. Konopka, "A maximum entropy principle for the distribution of local complexity in naturally occurring nucleotide sequences," *Computers & Chemistry*, vol. 16, no. 2, pp. 117–124, 1992.

[4] A. Milosavljević and J. Jurka, "Discovering simple DNA sequences by the algorithmic significance method," *Computer Applications in the Biosciences*, vol. 9, pp. 407–411, 1993.

[5] L. Allison, L. Stern, T. Edgoose, and T. I. Dix, "Sequence complexity for biological sequence analysis," *Computers & Chemistry*, vol. 24, pp. 43–55, 2000.

[6] O. Delgrange and E. Rivals, "STAR: an algorithm to search for tandem approximate repeats," *Bioinformatics*, vol. 20, no. 16, pp. 2812–2820, 2004.

[7] T. I. Dix, D. R. Powell, L. Allison, J. Bernal, S. Jaeger, and L. Stern, "Comparative analysis of long DNA sequences by per element information content using different contexts," *BMC Bioinformatics*, vol. 8, no. Suppl. 2, pp. S10, 2007.

[8] G. Korodi and I. Tabus, "An efficient normalized maximum likelihood algorithm for DNA sequence compression," *ACM Trans. on Information Systems*, vol. 23, no. 1, pp. 3–34, Jan. 2005.

[9] B. Behzadi and F. Le Fessant, "DNA compression challenge revisited," in *Combinatorial Pattern Matching: Proc. of CPM-2005*, Jeju Island, Korea, June 2005, vol. 3537 of *LNCS*, pp. 190–200, Springer-Verlag.

[10] G. Korodi and I. Tabus, "Normalized maximum likelihood model of order-1 for the compression of DNA sequences," in *Proc. of the Data Compression Conf., DCC-2007*, Snowbird, Utah, Mar. 2007, pp. 33–42.

[11] 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*, Snowbird, Utah, Mar. 2007, pp. 43–52.

[12] L. Allison, T. Edgoose, and T. I. Dix, "Compression of strings with approximate repeats," in *Proc. of Intelligent Systems in Molecular Biology, ISMB-98*, Montreal, Canada, 1998, pp. 8–16.

[13] D. Pratas and A. J. Pinho, "Compressing the human genome using exclusively Markov models," in *Advances in Intelligent and Soft Computing, Proc. of the 5th Int. Conf. on Practical Applications of Computational Biology & Bioinformatics, PACBB 2011*, Apr. 2011, vol. 93, pp. 213–220.

[14] A. J. Pinho, D. Pratas, and P. J. S. G. Ferreira, "Bacteria dna sequence compression using a mixture of finite-context models," in *Proc. of the IEEE Workshop on Statistical Signal Processing*, Nice, France, June 2011.

[15] A. J. Pinho, P. J. S. G. Ferreira, A. J. R. Neves, and C. A. C. Bastos, "On the representability of complete genomes by multiple competing finite-context (Markov) models," *PLoS ONE*, vol. in press, 2011.

[16] F. Nan and D. Adjeroh, "On the complexity measures for biological sequences," in *Proc. of the IEEE Computational Systems Bioinformatics Conference, CSB-2004*, Stanford, CA, Aug. 2004.

[17] L. Pirhaji, M. Kargar, A. Sheari, H. Poormohammadi, M. Sadeghi, H. Pezeshk, and C. Eslahchi, "The performances of the chi-square test and complexity measures for signal recognition in biological sequences," *Journal of Theoretical Biology*, vol. 251, no. 2, pp. 380–387, 2008.

[18] V. D. Gusev, L. A. Nemytikova, and N. A. Chuzhanova, "On the complexity measures of genetic sequences," *Bioinformatics*, vol. 15, no. 12, pp. 994–999, 1999.

[19] Y. L. Orlov and V. N. Potapov, "Complexity: an internet resource for analysis of DNA sequence complexity," *Nucleic Acids Research*, vol. 32, pp. W628–W633, 2004.

[20] A. J. Pinho, S. P. Garcia, P. J. S. G. Ferreira, V. Afreixo, C. A. C. Bastos, A. J. R. Neves, and J. M. O. S. Rodrigues, "Exploring homology using the concept of three-state entropy vector," in *Pattern Recognition in Bioinformatics, 5th IAPR Int. Conf., PRIB 2010*, Sept. 2010, vol. LNBI 6282, pp. 161–170.

Annex

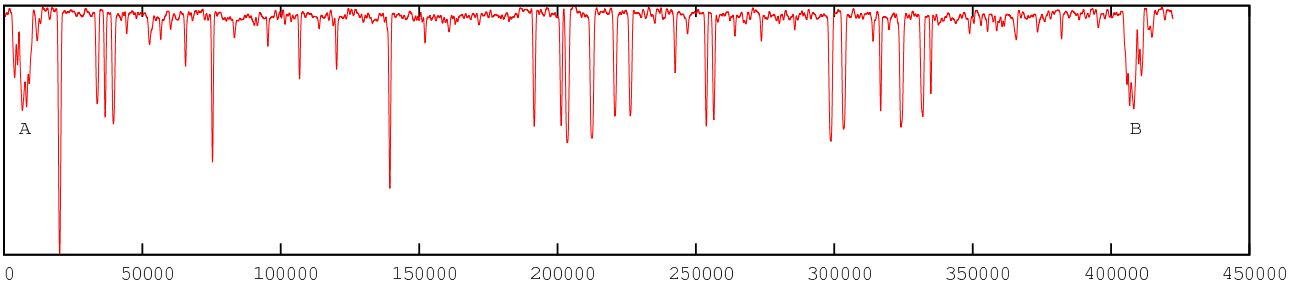

Figure 1: Preliminary complexity profile (relying only on the estimated code length) of chromosome 1 of *Cyanidioschyzon merolae*'s genome, generated by finite-context models. The horizontal axis indicates position along the chromosome. It is clear that some regions present lower complexity. In this project, we will investigate how this relates to biologically relevant features.

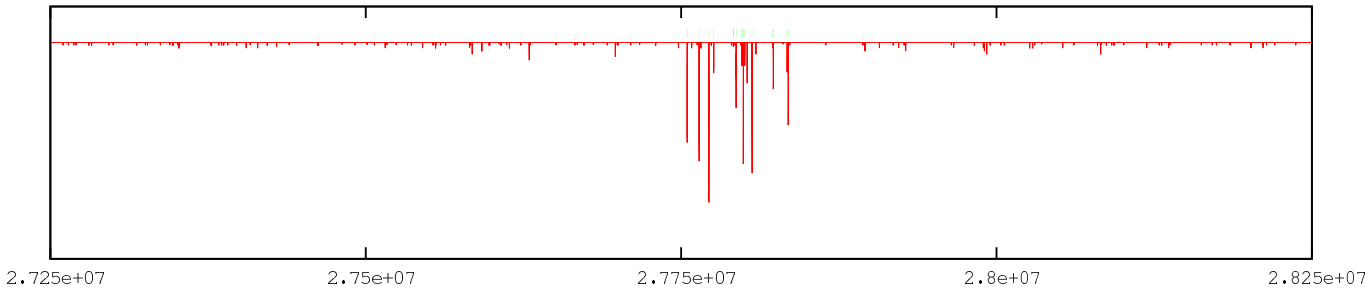

Figure 2: Preliminary complexity profile (relying only on the estimated code length) over a 1Mb window of human (*Homo sapiens*) chromosome 7. The green dashes match exon positions of gene TAX1BP1 (human T-cell leukemia virus type I binding protein 1), for comparison with the low-complexity peaks. The profile was generated by scanning the human chromosome with a finite-context model obtained from the homologous mouse (*Mus musculus*) gene. In this project, we will explore this property for establishing distances among the species.