Analysis of patterns in *S. pombe* genome through compression-based complexity profiles

Diogo Pratas
pratas@ua.pt

Sara P. Garcia
spgarcia@ua.pt

Armando J. Pinho
ap@ua.pt

Signal Processing Lab, IEETA / DETI,
University of Aveiro,
3810–193 Aveiro, Portugal

Abstract

It has been shown that finite-context (Markov) models are a powerful tool for representing DNA sequences, as demonstrated by the good compression results that they have been able to provide in recent works. However, they may also be useful in other tasks, such as in data analysis.

In this paper, we explore this line, studding the structural patterns in the *Schizosaccharomyces pombe* genome. Moreover, we analyse inter-chromosomal structural relations using complexity profiles. These complexity profiles allow a quick analysis, unveiling locations of low information content, which are usually associated with DNA regions of potential biological interest.

1 Introduction

In the context of DNA data compression, finite-context models have been used for describing the data in an efficient way, as demonstrated by the good compression results that they have been able to provide in recent works [2, 3, 4].

Finite-context models assume that the source has Markovian properties, i.e., that the probability of the next outcome of the information source depends only on some finite number of (recent) past outcomes. This past is normally referred to as the "context", hence the name "finite-context model".

The most obvious application of finite-context models is for data storage, i.e., compression. However, finite-context models can also be used in other approaches, such as in DNA data analysis, has we have shown in recent work[2]. This new direction suggests the exploration of DNA sequenced species, in order to unveil and understand biological designs and functions.

In this paper, we use finite-context models to create complexity profiles. Basically, a complexity profile indicates how many bits it is required to represent each symbol (DNA base). These complexity profiles are of interest because they reveal structures inside the chromosomes, structures that are often associated with regulatory functions of DNA.

Thereby, using complexity profiles, we study the genome of *Schizosaccharomyces pombe*, with the aim of searching biological important patterns. The *S. pombe*, also called "fission yeast", is a species of yeast composed by three pairs of chromosomes, totaling approximately 14.1 million base pairs.

2 Results and Discussion

In this study, we used the *S. pombe* genome (uid 127), obtained from the national Center for Biotechnology Information (NCBI), `ftp://ftp.ncbi.nlm.nih.gov/genomes/`.

In recent works, we have introduced a method [1], based on finite-context models, to build complexity profiles. In this paper, we analyse the *S. pombe* genome with the mentioned method.

Accordingly, we have extracted the complexity profiles for each of the three chromosomes. These results are presented in Figure 2. As it can be seen, there are locations of low information content which are clearly associated with DNA regions of biological interest, such as telomeric and centromere regions. Therefore, we have marked with letters A, C, D, F, G and I the telomeric regions and with letters B, E and H the centromere regions. Yet, these marked letters clearly identifies what is the long arm (q) and short arm (p) on each chromosome.

In some species, the centromeres are regions hard to find, specially when the low-pass filter has smaller bandwidth, due to the size of the sequence. However, as Wood *et al.* [5] investigated, the *S. pombe* centromeres are large comparing to the budding yeast *S. cerevisiae*. In this way, we could easily identify them with complexity profiles (B, E and H). Moreover, as it can be seen in Figure 1, the sizes of the centromeres regions vary inversely with the lengths of the chromosomes.

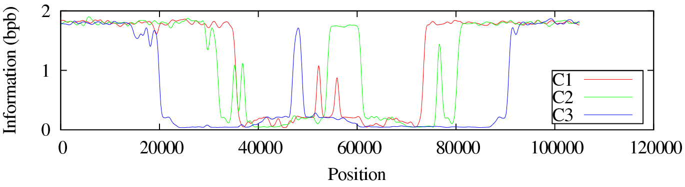

Figure 1: Plot of the information content of the centromeres from chromosome 1 (C1), 2 (C2) and 3 (C3). The information content was processed in both directions, combined using the minimum value of each direction, and low-pass filtered using a Blackman window of 1001 (drop: 20 bases).

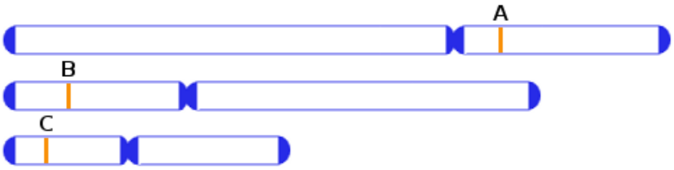

Figure 4: Illustration of the three chromosomes of *S. pombe* genome marked with genes ef1a-b (A), ef1a-c (B) and ef1a-a (C).

According to Wood *et al.* [5], possibly more extended centromeric regions are required for proper mitotic and meitotic behaviour when the chromosome arms are shorter.

Thereafter, we made an inter-chromosomal study in *S. pombe* genome. For this purpose, we resort to chromosome concatenation.

We have concatenated chromosome 1 with chromosome 3 and ran the compression (in both directions), presenting only the complexity profile of chromosome 3 (taking into account the statistics of chromosome 1), as it can be seen in Figure 3. The same process has been done for chromosome 1 and chromosome 2, however, due to space restrictions we only show plots for chromosome 3.

In Figure 3, we have unveiled important regions marked with the letters A, B and C. Starting with the letter B, this region contains the 2529 bases of gene eft202 (from base 537326 to 539854, in chr. 3). According, the statistics that unveiled this gene were extracted also from 2529 bases of chromosome 1, which represent gene eft201 (from base 2907701 to 2910229, chr. 1, with ∼99% sequence similarity to gene eft202).

In relation to the region marked with letter A (Figure 3), we have verified that chromosome 1 unveiled a repetition in chromosome 3, that representes two highly similar genes (ef1a-a in chromosome 3 and ef1a-b in chromosome 1 with ∼98% sequence similarity). Moreover, chromosome 2 unveild another very similar gene, ef1a-c, with ∼98% sequence similarity to both previous genes. In Figure 4, there is an illustration that shows the relative position of these genes. In this ilustration, letter A marks a region from base 4095202 to 4096584 (1383 bases, chr. 1). Letter B refers to base 626106 to 627488 (1383 bases, chr. 2), and letter C from base 268097 to 269479 (1383 bases, chr 3).

Interestingly, these very similar genes, present in all *S. pombe* chromosomes (a very rare property) and always in the short arm, have homologous in the following species: human, chimpanzee, dog, cow, rat, chicken, zebrafish, fruit fly, mosquito, *C. elegans*, *S. cerevisiae*, *K. lactis*, *E. gossypii*, *M. grisea* and *N. crassa*.

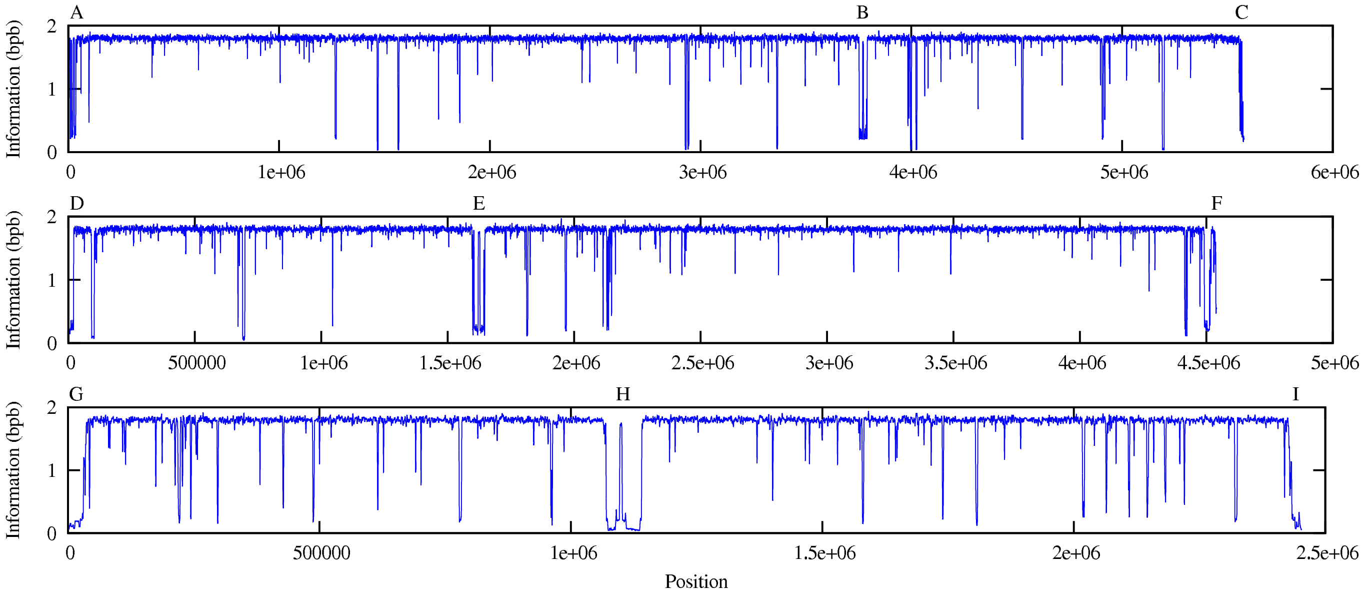

Figure 2: Plots of the information content for chromosome 1 (first row), chromosome 2 (second row) and chromosome 3 (third row) of *S. pombe*. The information content was processed in both directions, combined using the minimum value of each direction, and low-pass filtered using and a Blackman window of 1001 (drop: 20 bases).

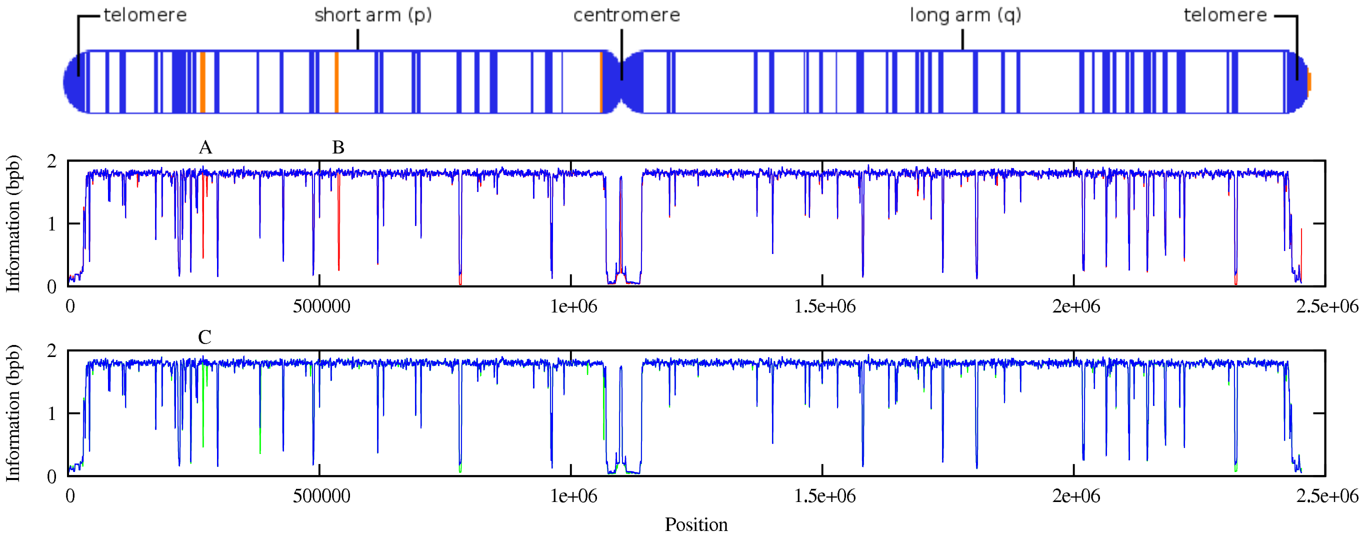

Figure 3: Information content for chromosome 3 of *S. pombe*. The first row shows a representation for chromosome 3 and their long repetitive zones. The second row shows chromosome 3 (blue) with information added from chromosome 1 (green). The third row shows chromosome 3 (blue) with information added from chromosome 2 (red). The information content was processed in both directions, combined using the minimum value of each direction, and low-pass filtered using a Blackman window of 1001 (drop: 20 bases).

3 Conclusions

Using finite-context models, we have unveiled locations of low information content, associated with DNA regions of biological interest in *Schizosaccharomyces pombe* genome, such as telomeric and centromere regions. Moreover, we have identified homologous genes present in more than one chromosome. As for example, the ef1a-derived genes which are contained in all chromosomes of this species.

We believe that this work will be a starting point for other intra-species analysis and furthermore for inter-species analysis (especially in larger genomes)

4 Acknowledgements

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-EIA/103099/2008). Sara P. Garcia acknowledges funding from the European Social Fund and the Portuguese Ministry of Science, Technology and Higher Education.

References

[1] A. J. Pinho and P. J. S. G. Ferreira. Finding unknown repeated patterns in images. In *Proc. of the 19th European Signal Processing Conf., EUSIPCO-2011*, Barcelona, Spain, August 2011.

[2] 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*, 6(6):e21588, 2011. doi: 10.1371/journal.pone.0021588.

[3] 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.

[4] 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*, volume 93, pages 213–220, April 2011.

[5] V. Wood et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature*, 415(6874):871–80, February 2002. doi: 10.1038/nature724.