Exon: a web-based software toolkit for DNA sequence analysis

Diogo Pratas, Armando J. Pinho, and Sara P. Garcia

Abstract Recent advances in DNA sequencing methodologies have caused an exponential growth of publicly available genomic sequence data. By consequence, many computational biologists have intensified studies in order to understand the content of these sequences and, in some cases, to search for association to disease. However, the lack of public available tools is an issue, specially when related to efficiency and usability. In this paper, we present Exon, a user-friendly solution containing tools for online analysis of DNA sequences through compression based profiles.

Key words: web-based software toolkit, DNA sequence analysis, DNA compression

1 Introduction

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 [7, 11]. For example, it has been observed that low complexity regions of DNA are often associated with important regulatory functions [6].

Several measures have also been proposed for evaluating the complexity of DNA sequences. 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) [4, 1, 5, 15].

One of the key advantages of DNA compression based on finite-context models [13, 10, 8, 9] is that the encoders are fast and have $\mathcal{O}(n)$ time com-

Signal Processing Lab, IEETA / DETI, University of Aveiro, 3810–193 Aveiro, Portugal
e-mail: {pratas,ap,spgarcia}@ua.pt

plexity. 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 proven to give good returns in terms of compression gains, but it may be disadvantageous in terms of time consuming to perform the compression. Although slow encoders could be tolerated for storage purposes (compression could be ran in batch mode), for interactive applications they are certainly not appropriate. For example, the currently best performing DNA compression technique, eXpert-Model (XM) [3], could take hours for compressing a single human chromosome. Compressing one of the largest human chromosomes with the techniques based on finite-context models (FCM) takes less than ten minutes. Along with this inconvenience, there is the need for a strong computational system to perform these operations, particularly with regard to memory (RAM). On the other hand, these tools require local installation, sometimes on particular operating systems, which makes them inconvenient to use. Moreover, there is the need of designing a graphical interface to make Exon attractive to biologists and not only informatics, instead of the prior command line approach.

In this paper, we provide solutions to the mentioned issues using Exon, a web-based user-friendly software toolkit that analyses DNA sequences using compression based approaches, such as finite-context modelling [13, 10, 9] and XM [3]. Since the software is web-based, it is available to any computer (with a web browser linked to the Internet) and without the need of any software installation. Moreover, Exon is hosted at a virtual web server composed by 8 cores of Intel(R) Xeon(R) CPU X5650 @ 2.67GHz, with at least 8 GB of RAM and with 2 TB of storage, running the CentOS linux distribution.

This paper is organized as follows. In Section 2, we describe the methods used, in particular the compression approaches. In Section 3, we provide some examples of the Exon usage. Finally, in Section 4, we draw some conclusions.

2 Materials and methods

Using technologies and programming languages such as HTML, PHP, Java, Javascript, CSS, PostgresSQL, shell script and C, we were able to integrate several methods in the Exon toolkit. All these methods fall into one of three categories (pre-encoding, encoding and post-encoding).

The first category is pre-encoding (sequence edition before encoding). In this category, there are the following methods: reverse (a sequence), concatenate (two sequences) and generate (a sequence). The first two methods are self-explanatory, due to their names. Sequence generation is based on multiple competing finite-context models. In short, this generator allows to generate a synthetic sequence based on statistics collected from a template sequence, using a stochastic process [12].

The second category is encoding (sequence compression). In this category, there are two possible models: XM or FCM (finite-context modelling). The first method, XM [3], relies on a mixture of experts for providing symbol by symbol probability estimates, which are then used for driving an arithmetic encoder. The algorithm comprises three types of experts: (1) order-2 Markov models; (2) order-1 context Markov models, i.e., Markov models that use statistical information only of a recent past (typically, the 512 previous symbols); (3) the copy expert, that considers the next symbol as part of a copied region from a particular offset. The probability estimates provided by the set of experts are then combined using Bayesian averaging and sent to the arithmetic encoder. Although the XM approach is inappropriate for large DNA sequences, we have included it in the Exon toolkit, because it can be used in small sequences (normally below 2 MB) and also for comparison with other methods. The other method, FCM [9], is an approach based on multiple finite-context models of different orders that compete for encoding the data. Figure 1 gives an example of these multiple models. The competitive procedure implies that the best of the models is chosen for encoding each DNA block, i.e., the one that requires less bits is used for representing the current block.

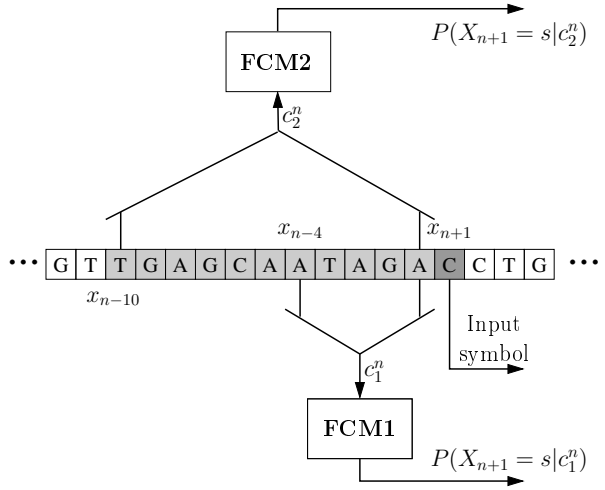

Fig. 1 Example of the use of multiple finite-context models for encoding DNA sequence data. In this case, two models are used, one with a depth-5 context and the other using an order-11 context.

The third (and last) category is the post-encoding (computation and manipulation of the information content received from category two). In this particular category, there is a module capable of processing the information content, by applying signal processing techniques, such as filtering, in order to improve the visual output of the complexity profiles.

The process of building a complexity profile for visualization can be seen in Figure 2.

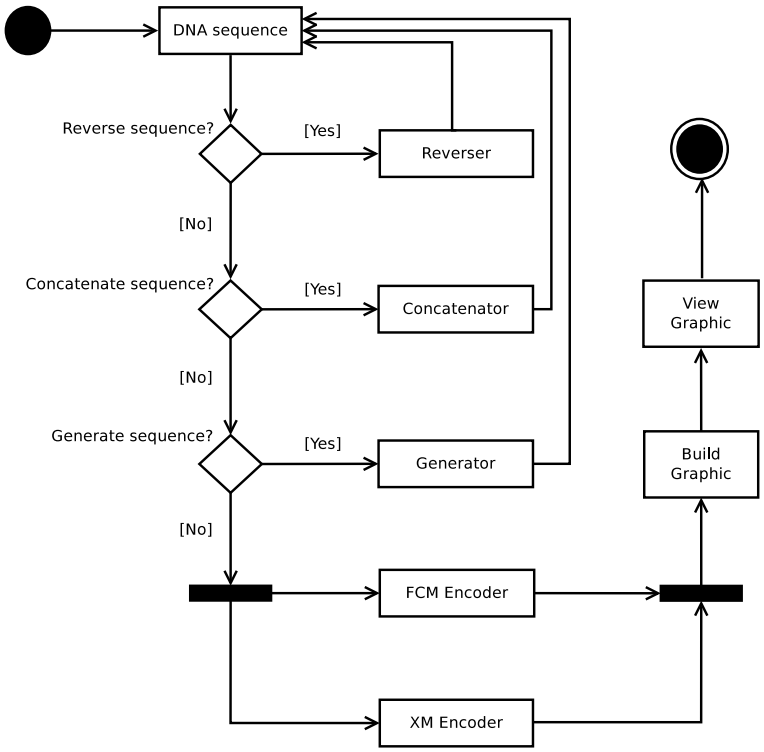

Fig. 2 Activity diagram of the process of building a complexity profile for visualization.

Exon has a pyramidal hierarchy structure of 3 levels (roles). The first and lowest is the guest user. Here, the user can perform the basic operations, namely those described in Figure 2, although, with certain restrictions such as limited number of FCM models to be chosen. In the middle, there is the operator level. This is a registered user that has the same privileges as the guest user, but with access to more features such as the uploading of sequences. The upper level is administrator. This is the unlimited user, that can also control logs, edit files, edit accounts, edit models, among other operations.

2.1 Software availability

This web-based software toolkit is publicly available for non-commercial purposes at http://exon.ieeta.pt.

3 Some examples

In this section, we provide some examples where we used Exon. For that purpose, we used the *Saccharomyces cerevisiae* genome (uid128) from the National Center for Biotechnology Information (NCBI).

Figure 3 shows an example of a complexity profile (corresponding to chromosome 2 of the *S. cerevisiae*), generated by a multiple finite-context model DNA encoder in Exon.

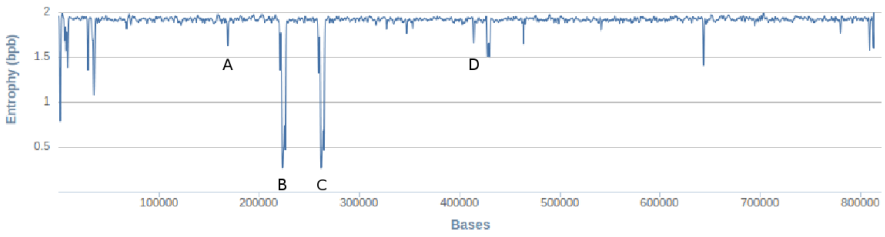

Fig. 3 Complexity profile of chromosome 2 from the *S. cerevisiae* genome. The information content was processed in both directions, combined using the minimum value of each direction, and low-pass filtered using a Blackman window of size 11.

We can observe several regions where the complexity is very small, meaning that a reduced number of bits was required for compressing those regions. Of particular interest are the two regions that are below 0.5 bpb (marked with letters B and C). These regions, which have been zoomed-in in Figure 4, correspond to transposons (sequences of DNA that can move or transpose themselves to new positions within the genome of a single cell). There are strong evidences that genetic diseases are caused by transposition events, for more information see [14].

The transposon marked with letter B (YBLWTy1-1) has 5,915 bases (from base 221,037 to base 226,952). The transposon marked with letter C (YBRWTy1-2) has 5,917 bases (from base 259,578 to base 265,494). Moreover, a Blast search [2] indicates that these sequences have ∼99% of identity.

Similarly, the regions marked with letters A and D represent homologous genes. In particular, the region marked with letter A, with 954 bases (from base 168,423 to base 169,376), represents the gene RPL19B. On the other hand, the region marked with letter D, with 1,076 bases (from base 414,186

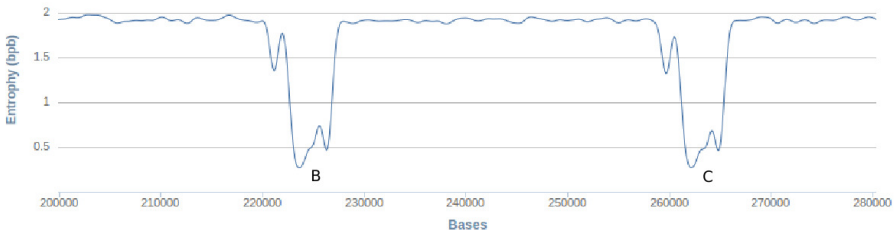

Fig. 4 Zoom of the complexity profile (from base 200,000 to base 280,000) of chromosome 2 from the *S. cerevisiae* genome. The information content was processed in both directions, combined using the minimum value of each direction, and low-pass filtered using a Blackman window of size 11.

to base 415,261), represents the gene RPL19A. A Blast search indicates that these sequences have ∼93% of identity.

Another application of Exon is the ability to perform inter-sequence analysis, i.e., using sequence concatenation it is possible to unveil low complexity zones which are usually associated to zones of potential biological interest. Thereby, we have concatenated chromosome 2 and 4 of of *S. cerevisiae* genome (showing just the results for the range of chromosome 2) and built the correspondent complexity profile (see Figure 5).

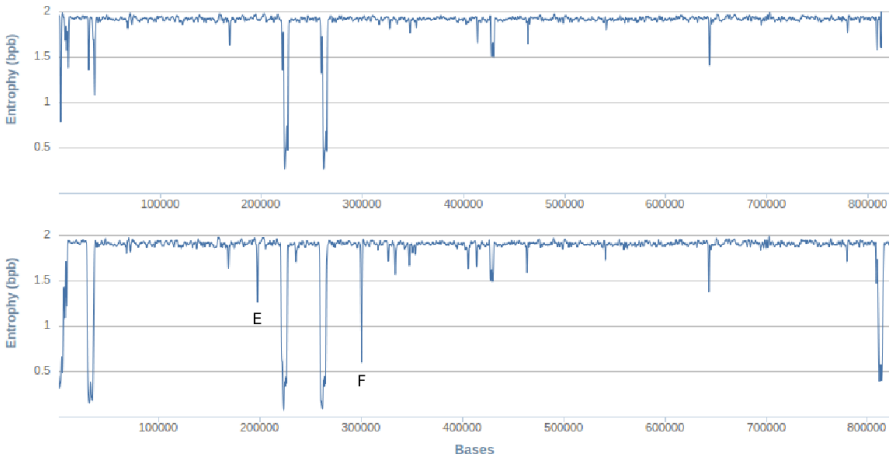

Fig. 5 Complexity profile comparison of chromosome 2 from the *S. cerevisiae* genome. The first row shows the information content of chromosome 2. The second row shows chromosome 2 with information added from chromosome 4. The information content was processed in both directions, combined using the minimum value of each direction, and low-pass filtered using a Blackman window of size 11.

In Figure 5, it is possible to observe (second row) that some new low complexity zones have been unveiled, comparing with the complexity profile of chromosome 2 alone (first row). These regions have been marked with letters E and F. Relatively to region E, the complexity profile shows two locations, almost coincident, of tRNA (tRNA-Ile and tRNA-Gly), which are contained also in chromosome 4. Letter F marks a region containing 1,089 bases, corresponding to gene RPL4A (from base 300,166 to base 301,254). Thereafter, we determined the corresponding source, leading to a region in chromosome 4 containing also 1,089 bases corresponding to gene RPL4B (from base 414,186 to base 415,261). Blast search indicates that these sequences have ∼100% of identity, i.e., these sequences are equal. These very similar genes have homologous in other species, such as in the *Homo sapiens*.

4 Conclusion

In this paper, we have presented Exon, a user-friendly solution containing tools for online analysis of DNA sequences through compression based profiles.

The main aim was to bridge the field of biology and the field of informatics, allowing a biologist without expertise in computer science to create complexity profiles and analyse DNA sequences. To attain that, we have provided a graphical interface, avoiding the command line approach, together with the availability of a powerful web server, not requiring software installation (with common usage, accessible to any computer connected to the Internet).

Also, we have demonstrated the usefulness of Exon, building complexity profiles of a few sequences from the *Saccharomyces cerevisiae* genome. In particular, we have identified transposons, tRNA genes, similar and highly similar genes.

5 Acknowledgements

This work was partially funded by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology in the context of the project FCOMP-01-0124-FEDER-010099 (FCT reference PTDC/EIA-EIA/103099/2008). Sara P. Garcia acknowledges funding from the European Social Fund and the Portuguese Ministry of Education and Science.

References

1. Allison, L., Stern, L., Edgoose, T., Dix, T.I.: Sequence complexity for biological sequence analysis. Computers & Chemistry **24**, 43–55 (2000)
2. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (1990). DOI 10.1006/jmbi.1990.9999
3. Cao, M.D., Dix, T.I., Allison, L., Mears, C.: A simple statistical algorithm for biological sequence compression. In: Proc. of the Data Compression Conf., DCC-2007, pp. 43–52. Snowbird, Utah (2007)
4. Crochemore, M., Vérin, R.: Zones of low entropy in genomic sequences. Computers & Chemistry pp. 275–282 (1999)
5. Dix, T.I., Powell, D.R., Allison, L., Bernal, J., Jaeger, S., Stern, L.: Comparative analysis of long DNA sequences by per element information content using different contexts. BMC Bioinformatics **8**(Suppl. 2), S10 (2007). DOI 10.1186/1471-2105-8-S2-S10
6. Gusev, V.D., Nemytikova, L.A., Chuzhanova, N.A.: On the complexity measures of genetic sequences. Bioinformatics **15**(12), 994–999 (1999)
7. Nan, F., Adjeroh, D.: On the complexity measures for biological sequences. In: Proc. of the IEEE Computational Systems Bioinformatics Conference, CSB-2004. Stanford, CA (2004)
8. Pinho, A.J., Ferreira, P.J.S.G.: Finding unknown repeated patterns in images. In: Proc. of the 19th European Signal Processing Conf., EUSIPCO-2011. Barcelona, Spain (2011)
9. Pinho, A.J., Ferreira, P.J.S.G., Neves, A.J.R., Bastos, C.A.C.: On the representability of complete genomes by multiple competing finite-context (Markov) models. PLoS ONE **6**(6), e21,588 (2011). DOI 10.1371/journal.pone.0021588
10. Pinho, A.J., Pratas, D., Ferreira, P.J.S.G.: Bacteria DNA sequence compression using a mixture of finite-context models. In: Proc. of the IEEE Workshop on Statistical Signal Processing. Nice, France (2011)
11. Pirhaji, L., Kargar, M., Sheari, A., Poormohammadi, H., Sadeghi, M., Pezeshk, H., Eslahchi, C.: The performances of the chi-square test and complexity measures for signal recognition in biological sequences. Journal of Theoretical Biology **251**(2), 380–387 (2008)
12. Pratas, D., Bastos, C.A.C., Pinho, A.J., Neves, A.J.R., Matos, L.: DNA synthetic sequences generation using multiple competing Markov models. In: Proc. of the IEEE Workshop on Statistical Signal Processing. Nice, France (2011)
13. Pratas, D., Pinho, A.J.: 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, vol. 93, pp. 213–220 (2011)
14. Roy, A., Carroll, M., Kass, D., Nguyen, S., Salem, A., Batzer, M., Deininger, P.: Recently integrated human Alu repeats: finding needles in the haystack. Genetica **107**(1-3), 149–61 (1999)
15. Troyanskaya, O.G., Arbell, O., Koren, Y., Landau, G.M., Bolshoy, A.: Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. Bioinformatics **18**(5), 679–688 (2002)