

# Estimating topological entropy of biological sequences using a factor oracle\*

(Extended abstract)

Arnaud Lefebvre

ABISS – UMR CNRS 6037, University of Rouen  
76821 Mont Saint-Aignan CEDEX, FRANCE

and

Thierry Lacroq

ABISS – LIFAR, University of Rouen  
76821 Mont Saint-Aignan CEDEX, FRANCE

## ABSTRACT

This article presents a heuristic which estimates the topological entropy of biological sequences. This method is based on the factor oracle, a space economical structure, which allows to handle very large sequences such as entire chromosomes.

**Keywords:** topological entropy, biological sequences, factor oracle, heuristic.

## 1. INTRODUCTION

Biological sequences can be viewed as strings build on a four-letter alphabet for DNA and RNA and on a twenty-letter alphabet for proteins. The composition of biological sequences is not homogenous: some zones are highly repetitive and some contains a lot of information. Most of the time these zones have biological meanings. It is important to be able to locate them easily. Some techniques exist to do so (see [4], [5] or [3]).

We propose a new technique to locate in a very fast way such zones in very large sequences. For this we use the factor oracle of a sequence. The factor oracle is an indexing structure (see [2] and [1]). It is a very compact data structure first designed to do string matching but it has proved very efficient to find repetitions in biological sequences [6] and to perform data compression [7]. This article is organized as follows. Section 2 gives the definition of the oracle. Then section 3 presents the basic idea of the

method. Section 4 exhibits some results. Finally section 5 gives our conclusions.

## 2. THE FACTOR ORACLE

Let  $p = p[1..m]$  be a word of length  $|p| = m$  over an alphabet  $\Sigma$ . The set of all the words build over  $\Sigma$  is denoted  $\Sigma^*$ . Let  $\varepsilon$  be the empty word ( $|\varepsilon| = 0$ ). A word  $w \in \Sigma^*$  is a *factor* of  $p$  if and only if  $p$  can be written  $p = uwv$  with  $u, v \in \Sigma^*$ .

The factor oracle of a word  $p$  of length  $m$ , denoted by  $Oracle(p)$ , is an automaton with the following properties:

- it has  $m + 1$  states;
- it has within  $m$  and  $2m - 1$  transitions;
- it recognizes at least all the factors of  $p$ .

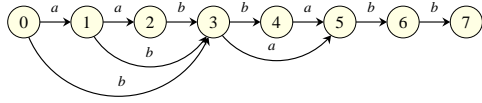
The factor oracle recognizes a bit more than all the factors of  $p$ . The exact characterization of its language is still an open question.

There is a bijection between the states of the oracle and the positions in  $p$  (state  $i$  corresponds to position  $i$  with  $1 \leq i \leq m$ , 0 is the start state). Each transition leading to state  $i$  is labeled by  $p[i]$ . We define two types of transitions: the transitions from state  $i$  to state  $i + 1$  ( $0 \leq i < m$ ) are called internal transitions and transitions from state  $i$  to state  $j$  such that  $j - i > 1$  are called external transitions. The oracle has exactly  $m$  internal transitions, thus to store it, one needs to store only at most  $m - 1$  external transitions without their labels. All the other informations can be deduced from the word  $p$ . An example is given figure 1.

\*This work was partially supported by a NATO grant PST.CLG.977017

This structure is linear in space, and its construction is linear in time (see proves in [2]).

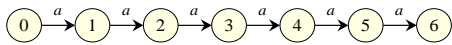
All this features enable the factor oracle to handle efficiently and in a fast way very large sequences.



**Figure 1:** Factor oracle of the word *aabbabb*. In addition to the word, the oracle consists in the transitions (0, 3), (1, 3) and (3, 5). All the other informations can be deduced from the word. All the states are terminal. The oracle recognizes all the factors of  $p$  and a bit more. For instance, *aba* is recognized though it is not a factor of *aabbabb*.

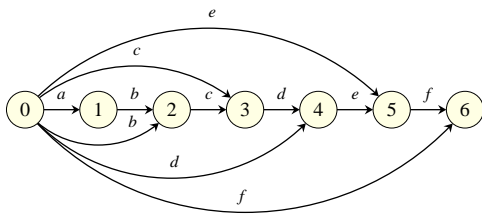
### 3. HEURISTIC FOR ESTIMATING THE TOPOLOGICAL ENTROPY

The number of external transitions in the factor oracle of a word  $p$  can give an idea on the number of different factors of  $p$  (and thus on the information conveyed by  $p$ ). For instance, the word  $a^m$  has exactly  $m + 1$  different factors (including the empty word) and its factor oracle has no external transition (which is the minimum possible).



**Figure 2:** Factor oracle of the word *aaaaaa*.

The word  $a_1a_2\dots a_m$  where  $\forall 1 \leq i, j \leq m, i \neq j, a_i \neq a_j$ , has exactly  $m \times (m + 1)/2 + 1$  different factors and its factor oracle has  $m - 1$  external transitions (which is the maximum possible).



**Figure 3:** Factor oracle of the word *abcdef*.

The word *aabbabb* has exactly 21 different factors and its factor oracle has 3 external transitions (see figure 1).

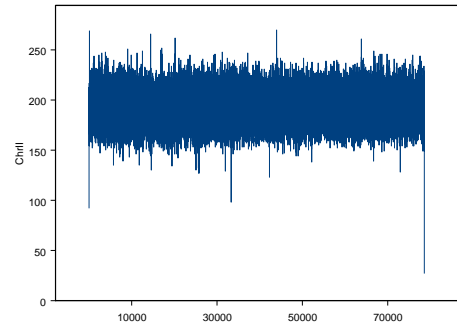
Thus, from these statements, we can conjecture the following:

**Conjecture:** the more different factors a word has, the more external transitions its factor oracle has.

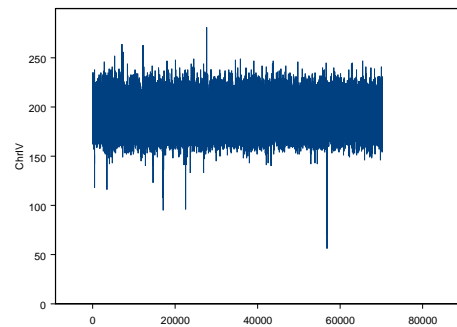
To estimate the validity of this conjecture in practical cases, we found areas of low or high topological entropy in long biological sequences such as chromosomes. For that, we partition the sequences into segments of the same length and we build the factor oracle for each segment. Then, for each oracle, we count the number of external transitions. We are thus able to distinguish segments of low or high entropy according to some thresholds on the number of external transitions.

### 4. EXPERIMENTAL RESULTS

Figures 4 to 8 show the results for chromosomes II and IV of *Arabidopsis thaliana*, chromosomes IV and IX of *Saccharomyces cerevisiae* and the chromosome of *Neisseria meningitidis* respectively. For each position multiple of 250, we computed the factor oracle of the segment composed of the 500 next bases. There is thus an overlap of 250 bases. For each oracle, we counted the number of external transitions. Each figure can be obtained in an automatic way from the raw nucleotide sequence in less than one minute per mega bp.

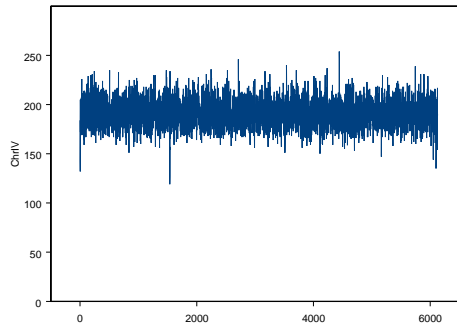


**Figure 4:** Chromosome II of *Arabidopsis thaliana*.



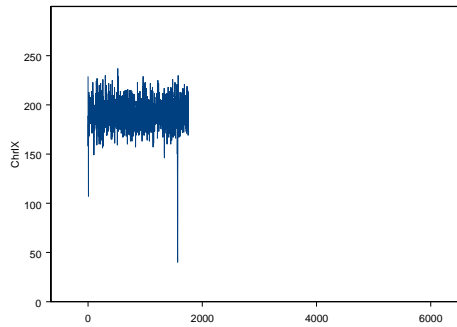
**Figure 5:** Chromosome IV of *Arabidopsis thaliana*.

The more clearly zones of low and high entropy of chromosomes II and IV of *Arabidopsis thaliana* are not described in genomic databanks.



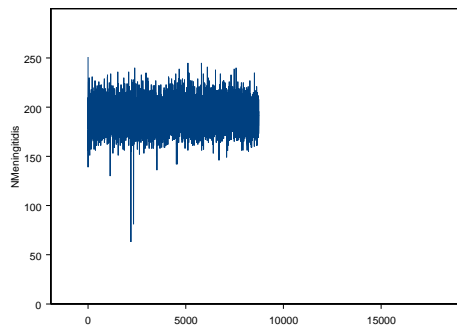
**Figure 6:** Chromosome IV of *Saccharomyces cerevisiae*.

The zone of low entropy appearing clearly at position 1539 in chromosome IV of *Saccharomyces cerevisiae* (see figure 6) corresponds, according to GenBank (<http://www.ncbi.nlm.nih.gov>), to an hypothetical protein.



**Figure 7:** Chromosome IX of *Saccharomyces cerevisiae*.

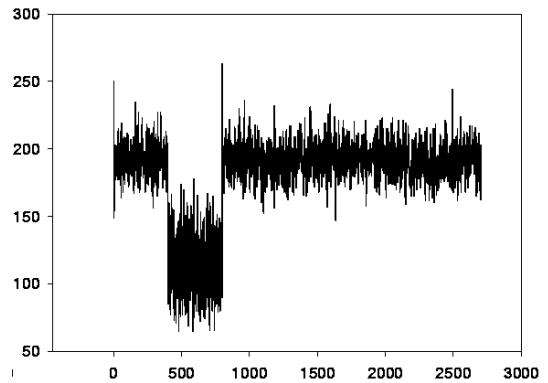
The zone of low entropy appearing clearly at position 1569 in chromosome IX of *Saccharomyces cerevisiae* (see figure 7) corresponds, according to GenBank, to a cell-wall protein. It was already found in [3] using a compact suffix automaton.



**Figure 8:** Chromosome of *Neisseria meningitidis*.  
The first zone of lowest entropy in the chromo-

some of *Neisseria meningitidis* corresponds, according to EMBL (<http://www.embl.org>), to a repeat unit of 465 base pairs direct repeat: it is a putative ferric enterobactin transporter binding protein. The second zone is a serie of repeat units.

Some experiments have been performed on chromosomes in which artificial segments have been inserted. Figure shows that an artificial insert of a segment built randomly on a two letter alphabet is clearly detected by our method.



**Figure 8:** An artificial insert in a chromosome is well detected by our method. This entropy measure has been obtained on chromosome V of *Saccharomyces cerevisiae*.

## 5. CONCLUSION AND PERSPECTIVES

We presented a new technique to locate zones of low or high topological entropy in DNA sequences. For that we compute the factor oracle on a window of length 500 for every position multiple of 250. To find long repetitions those parameters can be changed. We conjectured that the number of external transitions is related to the entropy of the segment in the window. This is confirmed in practice but it has to be formally proved. Our method is easy to implement and behave very well in practice. At this moment we rebuild each factor oracle from scratch but we conjecture that it is possible to build the factor oracle of  $p[i+k..j+k]$  from the factor oracle of  $p[i..j]$  in time  $O(k)$ .

## References

- [1] C. Allauzen. *Combinatoires sur les mots et recherche de motifs*. Ph.D. Thesis, University of Marne-la-Vallée, France, 2001.
- [2] C. Allauzen, M. Crochemore, and M. Raffinot. Factor oracle: a new structure for pattern matching. In J. Pavelka, G. Tel, and M. Bartosek, editors, *SOFSEM'99, Theory and Practice of Informatics*, number 1725, pp. 291–306, Milovy, Czech Republic, 1999.
- [3] M. Crochemore and R. Vérin. Zones of low entropy in genomic sequences. *Computers and Chemistry*, Vol. 23, No. 3–4, 1999, pp. 275–282.
- [4] M. Gribskov. The language metaphor in sequence analysis. *Computers Chemistry*. Vol. 16, No. 2, 1992, pp. 85–88.
- [5] A.K. Konopka. Theoretical Molecular Biology. In: *Encyclopaedia of Molecular Biology and Molecular Medicine*. VCII, Weinheim, Germany. Vol. 6, 1997, pp. 37–53.
- [6] A. Lefebvre and T. Lecroq. Computing repeated factors with a factor oracle. In L. Brankovic and J. Ryan, editors, *Proceedings of the 11th Australasian Workshop On Combinatorial Algorithms*, pp. 145–158, Hunter Valley, Australia, 2000.
- [7] A. Lefebvre and T. Lecroq. Compror: compression with a factor oracle. *Proceedings of Data Compression Conference*, 2001, pp. 502.