A Guaranteed Compression Scheme for Repetitive DNA Sequences*

E. Rivals†, J-P. Delahaye, M. Dauchet, O. Delgrange
Lab. d'Informatique Fondamentale de Lille
U.R.A. 369 du C.N.R.S., Université de LILLE I
Villeneuve d'Ascq FRANCE 59655

We present a text compression scheme dedicated to DNA sequences. The exponential growing of the number of sequences creates a real need of analyzing tools. A specific need emerges for methods that perform sequences classification upon various criteria, one of which is the sequence repetitiveness. A good lossless compression scheme is able to distinguish between "random" and "significative" repeats. Theoretical bases for this statement are found in Kolmogorov complexity theory.

In practice, classical dictionary compression schemes are unable to achieve compression on repetitive DNA sequences. Therefore, we conceived an optimized compression algorithm, called *Cfact*, which substitutes repeats by codes with pointers. The repeated factors may be very long and may have occurrences far away from each other.

Cfact executes sequentially the parsing phase and the encoding phase. To be sure to choose the most significant repeats, the **parsing phase** uses a suffix tree data structure to obtain the longest repeated factors. A repeat is significant if it can be encoded in order to obtain a local compression gain. The repeats are considered in decreasing length order. The **encoding phase** then considers the sequence from left to right. It encodes all non repetitive segments and first occurences of repetitive segments without compression, using two bits for each letter. The second occurrence of the repeated segment is replaced by a pointer (position and length) to the first occurence. Next occurences are replaced by the indexed identifier of the repeat.

If the length of the sequence is n then the time and space complexities of *Cfact* are in $O(n^2)$. Moreover, we prove that the compression is guaranteed if at least one "sufficiently" long repeat occurs in the sequence. A theorem gives a sufficient condition on this repeat length. The global compression gain is the sum of all local gains.

We experienced different compression schemes on long DNA sequences from different species, and compared their compression performances. The results and an extended abstract are available by FTP at *ftp.lifl.fr:/pub/projects/BIC/biologie/Cfact*. These experiments show that, on one hand, *Cfact*compresses repetitives sequences, while classical dictionary schemes do not. So, *Cfact* is able to find and encode DNA repetitions. On the other hand, many sequences especially from prokaryotic organisms give a null compression rate even for *Cfact* showing thanks to the compression guarantee, that a majority of these sequences are not repetitive. Therefore *Cfact* can effectively been employed for sequences classification purposes.

---

*Partially supported by the "GDR 1029 Informatique et Génomes" and by the "Groupement de Recherches et d'Études sur les Génomes"

†email: rivals@lifl.fr