Finite-context modeling of DNA sequences

Armando J Pinho,[1] António J R Neves,[1] Carlos A C Bastos,[1] and Paulo J S G Ferreira[1]
[1]*Signal Processing Lab, DETI / IEETA, University of Aveiro*

Motivation:

In the last fifteen years, several contributions have been made in the area of DNA data sequence compression. The proposed techniques share a common approach, which reflects the non-stationary nature of DNA data. They are typically composed of two methods, one based on Lempel-Ziv-like substitutional procedures, the other relying on low-order context-based (Markov) modeling. Usually, repetitive regions of the DNA sequences are represented using a pointer to a past occurrence of the repetition and the length of the repeating sequence. Both exact and approximate repetitions have been explored, as well as their inverted complements. When the substitutional method is not able to provide a satisfactory performance, then the region of the DNA sequence under evaluation is represented by a low-order finite-context model. One of the drawbacks of these substitutional approaches is the associated computational complexity. In fact, most of the effort spent by these encoding techniques is in the task of finding good exact/approximate repeats or inverted complements.

Background:

Low-order finite-context models have been used for DNA sequence compression as a secondary, fall back method. However, we have recently shown that models of orders higher than four are indeed able to attain significant performance. Moreover, we introduced new updating mechanisms that permit capturing information regarding the inverted repeats usually found in DNA sequences [1] and we investigated several aspects related to sets of finite-context models that compete for encoding the data [2].

Results:

The results obtained with the human genome show that an appropriate combination of finite-context models is able to represent well DNA sequence data, although not so well as methods also including substitutional approaches. Nevertheless, finite-context modeling may play an important role when low computational complexity is needed, such as in interactive applications requiring the computation of complexity profiles [3]. Moreover, these results also show that DNA sequence data may be described reasonably well by statistical models that rely only on the immediate past.

Conclusions:

Modeling DNA data using only finite-context models has advantages over the typical DNA compression approaches that mix purely statistical (for example, finite-context models) with substitutional models (such as Lempel-Ziv based algorithms): (1) finite-context models lead to much faster performance, a characteristic of paramount importance for long sequences; (2) the overall model might be easier to interpret, because it is made of sub-models of the same type.

References:

1. A. J. Pinho, A. J. R. Neves, and P. J. S. G. Ferreira. Inverted-repeats-aware finite-context models for DNA coding. *EUSIPCO-2008*, Lausanne, Switzerland, Aug 2008.

2. A. J. Pinho, A. J. R. Neves, C. A. C. Bastos, and P. J. S. G. Ferreira. DNA coding using finite-context models and arithmetic coding. *ICASSP-2009*, Taipei, Taiwan, Apr 2009.

3. 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*, 8(Suppl 2):S10, 2007.