PhD Proposal MAP-I (2013/14)

DNA sequence analysis with information profiles

Supervisors: Armando J Pinho (ap@ua.pt) — Paulo JSG Ferreira (pjf@ua.pt)

Research Unit: IEETA - Institute of Electronics and Informatics Engineering of Aveiro

University of Aveiro

Motivation

Data summarization and triage is one of the current top challenges in visual analytics. The goal is to let users visually inspect large data sets and examine or request data with particular characteristics. The need for summarization and visual analytics is also felt when dealing with digital representations of DNA sequences. Genomic data sets are growing rapidly, making their analysis increasingly more difficult, and raising the need for new, scalable tools. For example, being able to look at very large DNA sequences while immediately identifying potentially interesting regions would provide the biologist with a flexible exploratory and analytical tool.

Recently, we presented a new concept, the "information profile", which provides a quantitative measure of the local complexity of a DNA sequence, independently of the direction of processing. The computation of the information profiles is computationally tractable: we have shown that it can be done in time proportional to the length of the sequence.

Also, we have shown that information profiles are useful to detect large-scale genomic regularities by visual inspection. Several discovery strategies are possible, including the standalone analysis of single sequences, the comparative analysis of sequences from individuals from the same species, and the comparative analysis of sequences from different organisms. The comparison scale can be varied, allowing the users to zoom-in on specific details, or obtain a broad overview of a long segment.

Having this new concept at hand, we intend now fully explore its capabilities in a number of possible applications, such as the detection of homologous genes or the direct analysis of the data produced by high-throughput sequencing machines.

Bibliography

- Armando J Pinho, Sara P Garcia, Diogo Pratas, Paulo JSG Ferreira. *DNA sequences at a glance.* **PLoS ONE**, vol. 8, no. 11, p. e79922, November 2013 (DOI: 10.1371/journal.pone.0079922)
- Armando J Pinho, Diogo Pratas, Paulo JSG Ferreira, Sara P Garcia. *Symbolic to numerical conversion of DNA sequences using finite-context models.* Proc. of the European Signal Processing Conference, EUSIPCO 2011, Barcelona, Spain, p. 2024-2028, August 2011
- Armando J Pinho, Paulo JSG Ferreira, António JR Neves, Carlos AC Bastos. *On the representability of complete genomes by multiple competing finite-context (Markov) models.* **PLoS ONE**, vol. 6, no. 6, p. e21588, June 2011 (DOI: 10.1371/journal.pone.0021588)