PhD Proposal MAP-I (2013/14)

Compression of genomic data

Supervisor: Armando J Pinho (ap@ua.pt)

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

University of Aveiro

Motivation

Saying that the volume of genomic data produced every day is large is clearly an euphemism. In fact, with the dramatic drop in price of the sequencing machines (the one thousand dollar limit for sequencing a human genome will be history shortly), virtually everyone will want to sequence everything. Unfortunately, the pace at which storage and communication resources are evolving is not enough and the genomic data centers are being flooded with data. It is a data deluge.

For daily use, general purpose compression tools, such as gzip, may continue to play an important role in the context of genomics data processing, mainly due to its pervasiveness and relatively good speed. However, we have shown that special purpose compression tools can sometimes attain additional file reductions as large as 50% or even more, in relation to gzip. The possibility to virtually double the amount of sequence data that can be stored in a given space, exclusively by means of software compression tools, is an opportunity worthy of consideration by the genomics laboratories. Of course, higher compression can only be obtained using more complex algorithms, often requiring some more time and memory to run. However, these additional requirements are compensated by the relief attained in terms of storage requirements.

We intend to continue the work on the development of models and algorithms for compressing diverse types of genomic data, pushing forward the state of the art that we have been attaining. To achieve this, we are seeking highly motivated students, with good programming skills and the desire to be in the front line of the world research in this area.

Bibliography

- Armando J Pinho, Diogo Pratas. *MFCompress: a compression tool for FASTA and multi-FASTA data.* **Bioinformatics**, vol. 30, no. 1, p. 117-118, January 2014 (DOI: 10.1093/bioinformatics/btt594)

- Manuel J Duarte, Armando J Pinho. *Bacterial DNA sequence compression models using artificial neural networks.* **Entropy**, vol. 15, p. 3435-3448, August 2013 (DOI: 10.3390/e15093435)

- Luís Matos, Diogo Pratas, Armando J Pinho. *A compression model for DNA multiple sequence alignment blocks.* **IEEE Transactions on Information Theory**, vol. 59, no. 5, p. 3189-3198, May 2013 (DOI: 10.1109/TIT.2012.2236605)

- Armando J Pinho, Diogo Pratas, Sara P Garcia. *GReEn: a tool for efficient compression of genome resequencing data.* **Nucleic Acids Research**, vol. 40, no. 4, p. e27, February 2012 (DOI: 10.1093/nar/gkr1124)

- 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)