Information profiles for DNA pattern discovery

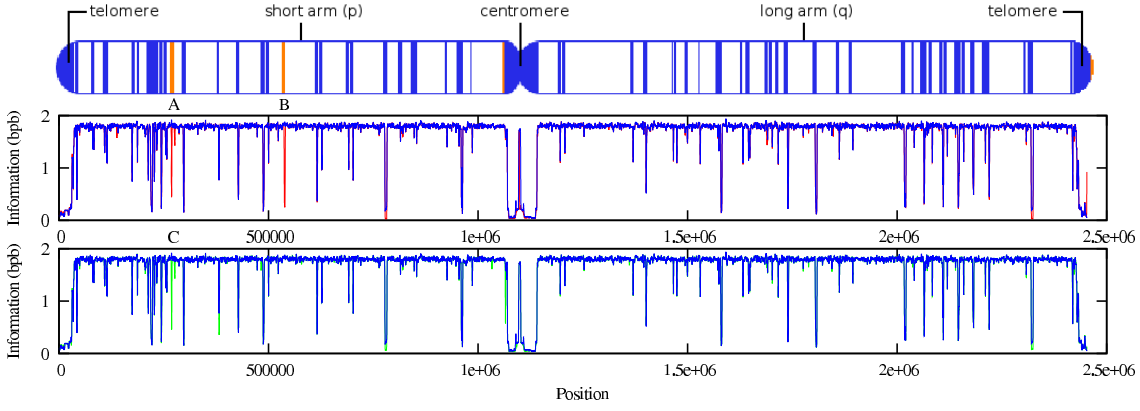
Armando J. Pinho, Diogo Pratas, and Paulo J. S. G. Ferreira

IEETA / Dept of Electronics, Telecommunications and Informatics
University of Aveiro, 3810–193 Aveiro, Portugal
ap@ua.pt — pratas@ua.pt — pjf@ua.pt

We describe an algorithm to detect genomic regularities within a *blind* discovery strategy. This algorithm uses information profiles built using an efficient DNA sequence compression method. The accurate matching of the low-information regions to annotated repetitive genomic structures, such as the centromeric and telomeric regions of a chromosome, suggests information profiles may be useful in *de novo* discovery of large-scale genomic regularities. Clearly, it is not possible to infer the genomic sequence *per se* from the information profiles, or the location of genomic regularities within base pair resolution. However, it is possible to discover the presence of regularities on a genome-wide scale, which may be useful for an exploratory genome analysis or for genome comparisons.

The algorithm relies on the efficient probabilistic modeling of the genomic sequence based on finite-context models and is sufficiently flexible and powerful to enable addressing biological questions and quickly obtaining the corresponding information profiles for a first-hand assessment. The creation of information profiles does not require high performance computational facilities. Building an information profile requires a computation time that depends only linearly on the size of the sequence (the information profile of a human chromosome can be created in a laptop computer in just a few minutes). The amount of computer memory required does not depend on the size of the sequence, but only on the depth of the finite context models used.

Below we show the information content of chr III of *S. pombe*. The first row shows a representation for chr III and their long repetitive zones. The second row shows chr III (blue) with information added from chr I (green). The third row shows chr III (blue) with information added from chr II (red).



Supplementary material in http://arxiv.org/abs/1401.4725.

Work supported in part by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology, in the context of the projects FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011) and Incentivo/EEI/UI0127/2013.

1068-0314/14 $31.00 © 2014 IEEE
DOI 10.1109/DCC.2014.54