A conditional compression distance that unveils insights of the genomic evolution

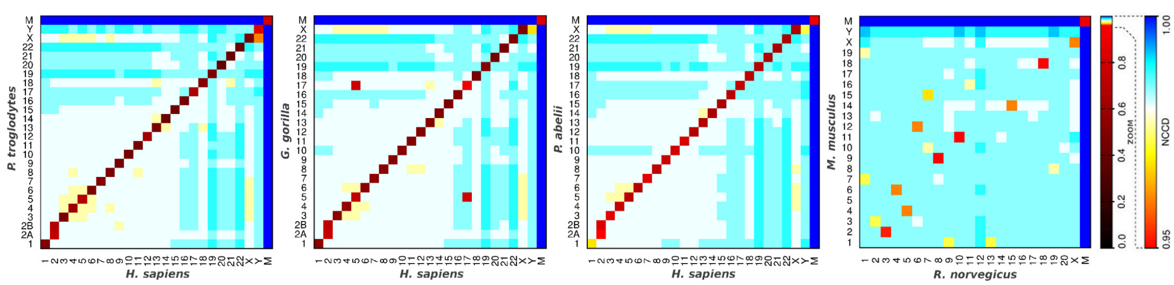
Diogo Pratas and Armando J. Pinho

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

We describe a compressed-based metric for measuring distances between genomic sequences. Instead of using the usual conjoint information content, as in the classical Normalized Compression Distance (NCD), it uses the conditional information content, denominated Normalized Conditional Compression Distance (NCCD), under a Kolomogorov complexity analogue, defined for objects x and y as

$$\mathrm{NCCD}(x,y) = \frac{\max\{C(x|y), C(y|x)\}}{\max\{C(x), C(y)\}},$$

where "Conditional" means that compressor C needs to be able to perform conditional compression. This approach requires a *normal* conditional compressor, that we define and assess in this work. The compressor is constituted by a set of multiple static and dynamic finite-context models, that cooperate under a supervision mixture model. It is able to handle several types of mutations (see supplementary material). We applied this metric to calculate chromosomal distances between *Hominidae* primates, and between *Muroidea* rodents (rat and mouse), as exemplified below.



For all primate species there is a direct correlation with the respective chromosome (C) number, with the exception of C2, justified by a fusion. Human CY is highly related with CX of other primate species, namely the *P. troglodytes*, because of CY / CX genetic information exchange. The *G. gorilla* and *H. sapiens* C5 / C17 show a translocation. *M. musculus* (MM) and *R. norvegicus* (RN) C18 and CX seem to be *diagonal* homologous. Subsequent analysis show other strong similarities, such as MM C4 / RN C5, MM C6 / RN C4, MM C12 / RN C6 and MM C14 / RN C15.

Common biological approaches (FISH) to unveill large-scale regularities are very time-consuming and expensive. This compressed-based metric allows to overcome these drawbacks, adding automation and possibility of parallelization.

Supplementary material in http://arxiv.org/abs/1401.4134.

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.2014.58