



Workshop on Modeling of Genetic Regulatory and Metabolic Networks

Valparaíso Complex Systems Institute -ISCV- <http://www.iscv.cl>
Valparaíso, Chile, ISCV- March 27th - 28th , 2008



Abstract

Alair Pereira do Lago

An efficient and lossless filter for multiple repeats with bounded edit distance

Similarity search in texts, notably in biological sequences, has received substantial attention in the last few years. Numerous filtration and indexing techniques have been developed in order to speed up the solution of the problem. However, previous filters were made for speeding up pattern matching, or for finding repeats between two strings or occurring twice in the same string. In this paper, we present an algorithm called *tuiuiu* for filtering strings prior to finding repeats under a bounded and quite large edit distance occurring twice or more in a string, or in two or more strings. Experimental results show that the filter can be very efficient: pre-processing with *tuiuiu* a 5.5 Mbp cross-species mammalian dataset where one wants to find functional elements using a multiple local alignment tool such as *Glam*, the overall execution time can be reduced from 11.5 hours to 20 minutes, bringing also improvements on the quality of the founded repeats. We have also applied *tuiuiu* to other different datasets, including a whole human chromosome.

www.iscv.cl