Abstract of Dissertation presented to UFF as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

Mutual information analysis of homologous DNA sequences

Luciana de Souza Pessôa

October/2004

Advisor: Helena Cristina da Gama Leitão
Co-advisor: Jorge Stolfi
Department: Computer Science

In this dissertation, we describe a method for estimating the amount of mutual information of homologous DNA sequences, i.e., the information contained in a DNA sequence that can be used to identify homologous blocks that have a same ancestor. For that purpose, we use signal processing techniques, especially spectral analysis, signal filtering, and information theory. The analysis of the mutual information content allows us to estimate the probability of false positives — strings that are not homologous to the given sequence, but are just as similar to it as the homologous ones.

In 1999, Leitão and Stolfi developed an efficient algorithm for the reconstruction of fragmented seramic objects, using the technique of multiscale sequence comparison. This technique may be applicable to the problem of finding similar strings in a bio-sequence database, which is a fundamental problem for the identification of homologous genes and for the assembly of genomes from sequenced fragments. The viability of multiscale comparison for this problem relies on the hypothesis that, even in the coarsest scales used, a DNA sequence still contains enough mutual information to eliminate a significative fraction of false positives. We verify this hipothesis by the method described here.