ENTROPIC PROFILER – efficient whole genome analysis using information theory and statistical concepts

Francisco Fernandes, Ana T. Freitas, Jonas S. Almeida, Susana Vinga

Entropic Profiles (EP) are local information plots that indicate overall conservation of motifs in genomes. They are based on Information Theory concepts and can be used as a method to extract and classify relevant and statistically significant segments of DNA sequences. The study of these motifs is very relevant because under or over-representation segments are often associated with significant biological meaning.

EP plots express the relative abundance of corresponding motifs for each position and allow estimating local relevant scales. Its calculation is based on the continuous Rényi quadratic entropy, using the Parzen window density estimation method applied to the Chaos Game Representation (CGR) of a sequence. For each position, the EP function retrieves information about the L-tuple suffixes directly from the density kernel function, which allows the extraction of scale independent motifs.

ENTROPIC PROFILER, the present tool implementation, is based on new data structures and algorithmic simplifications and allows to process whole genomes in just a few minutes. This major improvement is achieved by using k-truncated suffix trees, which have limited depth, side-links and shift-and algorithm. Furthermore, algebraic simplifications of the formulae used to calculate the profiles leads to a significant enhancement of the performance due to the avoidance of repeated calculations of mean and variance values.

Examples provided for *E. coli* and *H. influenzae* genomes evidence the retrieval of statistically and biologically relevant motifs, such as Chi sites (crossover hotspot instigator) and uptake signal sequences (USS+), respectively.

ENTROPIC PROFILER is freely available at http://kdbio.inesc-id.pt/software/ep/ as a web interface and as downloadable source code.