Whole genome analysis through Rényi Entropic Profiles

Susana Vinga[a,b], **Jonas S. Almeida**[c,d]

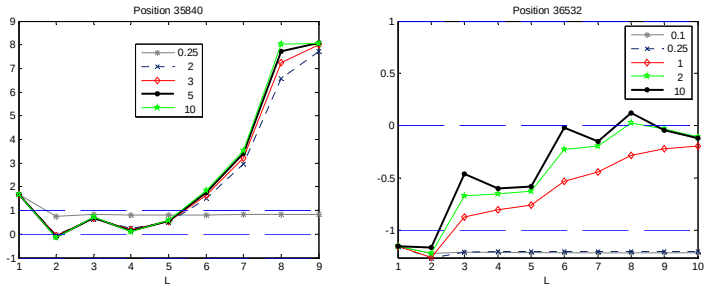
[a] *INESC-ID*, Portugal, [b] *FCM/UNL*, Portugal
[c] *Univ. Texas MDAnderson Cancer Center*, USA, [d] *ITQB/UNL*, Portugal

Genome sequences display overlapping signals on different scales, from single short DNA motifs to whole genes. The extraction and classifications of such information is still a significant challenge in computational biological sequence analysis.
Entropic profiles are local information plots for each position/symbol in a genome sequence. They can be obtained with iterative function systems for DNA by estimating point densities in Chaos Game Representation (CGR) maps, using Parzen's window estimation method coupled to a new fractal kernel function. Alternatively, they were shown to be obtainable also thought suffixes counts for each position (ranging from 1 to L-tuples) with two parameters ***L*** (memory/resolution which translates the Markov chain order) and ϕ, a smoothing parameter that weights differently the resolutions up to $L \geq 1$:

$$\hat{f}_{L,\phi}(x_i) = \frac{1 + \frac{1}{N}\sum_{k=1}^{L} 4^k \phi^k \cdot c([i-k+1,i])}{\sum_{k=0}^{L} \phi^k}$$

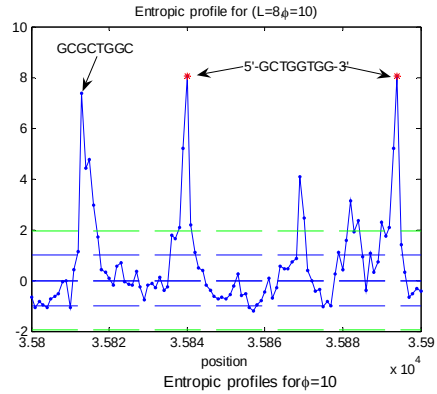
where $c([i-k+1,i])$ is the number of motifs $(s_{i-k+1} \cdots s_i)$ in the whole sequence of length *N*. After normalizing this value for all positions in the sequence, graphs with EP values can be obtained that express, for each position, the relative abundance of the corresponding motifs.

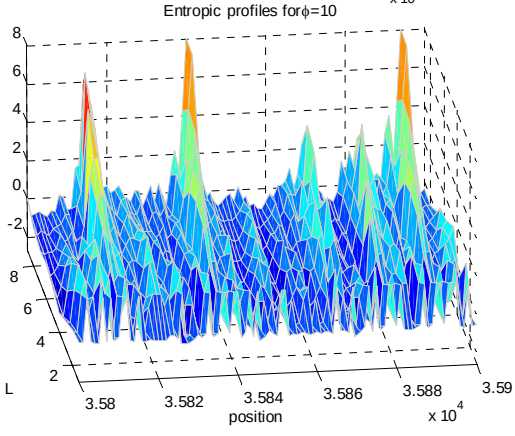
The detection of relevant and statistically significant segments can be accomplished unsupervisedly by spanning the parameters space to find local maxima.



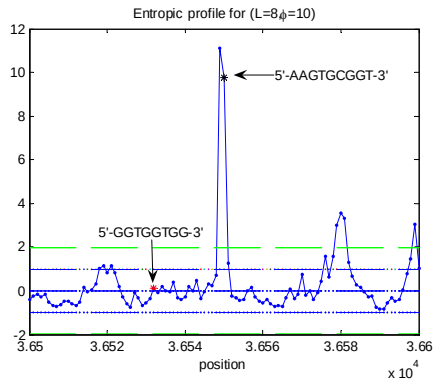
This application shows the detection of **Chi sites** (crossover hotspot instigator) in *Escherichia coli* K12 (`5′-GCTGGTGG-3′`) and **Uptake Signal Sequences** (USS+) in *Haemophilus influenzae* Rd (`5′-AAGTGCCGGT-3′`) genomes when processing their whole DNA, showing that the method correctly detected the corresponding scales and motifs present.

E. coli





H. influenzae



References
Almeida, J. S. and Vinga, S. *Algorithms Mol Biol.* 2006, **1**:18.
Vinga, S. and Almeida, J. S. *J Theor Biol* 2004, **231**:377-388.