Whole genome analysis through Rényi Entropic Profiles

Susana Vinga\textsuperscript{a,b}, Jonas S. Almeida\textsuperscript{c,d}

\textsuperscript{a} INESC-ID Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento, Portugal
\textsuperscript{b} FCM/UNL Faculdade de Ciências Médicas – Universidade Nova de Lisboa, Portugal
\textsuperscript{c} Univ. Texas MD Anderson Cancer Center, USA
\textsuperscript{d} ITQB/UNL Instituto de Tecnologia Química e Biológica – Universidade Nova de Lisboa, Portugal

Abstract
Rényi Entropic Profiles (EP) represent local information for each symbol in DNA sequences based on Information Theory. This methodology allows to infer automatically local scales and to detect exceptional suffixes, here illustrated for the analysis of E.coli and H.influenzae whole genomes, where Chi sites and Uptake Signal Sequences are correctly retrieved.

Introduction
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.

Methods
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.

Results
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 instigation) in Escherichia coli K12 (G\textsuperscript{+}C\textsuperscript{+} versus G\textsuperscript{−}C\textsuperscript{−}) and Uptake Signal Sequences (USSs) in Haemophilus influenzae Rd (\textsuperscript{a}H. influenzae\textsuperscript{b}) genomes when processing their whole DNA, showing that the method correctly detected the corresponding scales and motifs present.

Conclusions
\begin{itemize}
\item Entropic profiles (EP) provide useful local information about global features of DNA
\item Spanning EP parameter space for each position allows to find local extremes values with local scale interpretation
\item Detection of local scales is directly related with suffix and motifs over or under-representation and are correctly identified
\item Tests on whole genomes corroborate the strengths of this approach to detect biologically meaningful DNA segments
\end{itemize}

References
\begin{itemize}
\item Vinga, S. and Almeida, J.S. 2007 (submitted).
\end{itemize}

Acknowledgments
The authors acknowledge financial support by projects DynaMe (PTDC/EEA-AGR/69530/2006, FCT), MaGiC (IE02D01054, INESC-ID) and the Sixth Framework Programme of the European Union - Biosapiens project.