B59

Whole genome analysis through Rényi Entropic Profiles

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

[a] *INESC-ID Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento,* Portugal
[b] *FCM/UNL Faculdade de Ciências Médicas – Universidade Nova de Lisboa,* Portugal
[c] *Univ. Texas MDAnderson Cancer Center*, USA
[d] *ITQB/UNL Instituto de Tecnologia Química e Biológica – Universidade Nova de Lisboa*, Portugal

svinga@algos.inesc-id.pt
jalmeida@mdanderson.org

1 Abstract

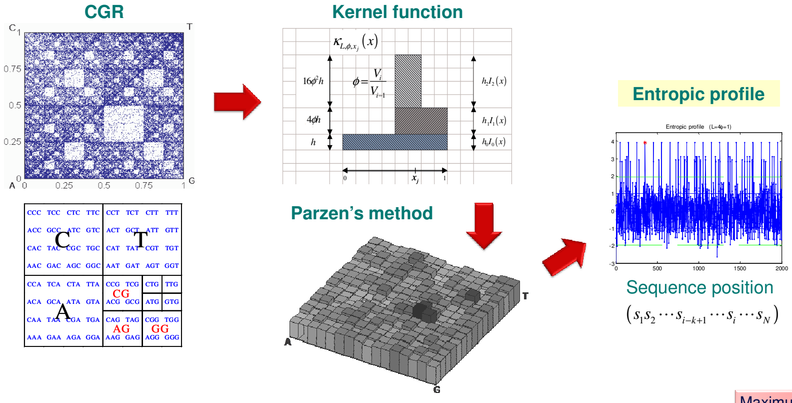
Rényi Entropic Profiles (EP) represent local information for each symbol in DNA sequences based on Information Theory. This methodology allows to infer automatically local scales and to detect exceptional suffixes, here illustrated for the analysis of *E.coli* and *H.influenza* whole genomes, where Chi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

2 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.

3 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.



CCC	TCC	CTC	TTC	CCT	TCT	CTT	TTT
ACC	GCC	ATC	GTC	ACT	GCT	ATT	GTT
CAC	TAC	CGC	TGC	CAT	TAT	CGT	TGT
AAC	GAC	AGC	GGC	AAT	GAT	AGT	GGT
CCA	TCA	CTA	TTA	CCG	TCG	CTG	TTG
ACA	GCA	ATA	GTA	ACG	GCG	ATG	GTG
CAA	TAA	CGA	TGA	CAG	TAG	CGG	TGG
AAA	GAA	AGA	GGA	AAG	GAG	AGG	GGG

C T A CG AG GG

$(s_1s_2\cdots s_{i-k+1}\cdots s_i\cdots s_N)$

Alternatively, entropic profiles 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 $L\geq1$

Entropic profile for the i^{th} symbol s_i, coordinate x_i

Number of motifs $(s_{i-k+1}\ldots s_i)$ in the whole sequence

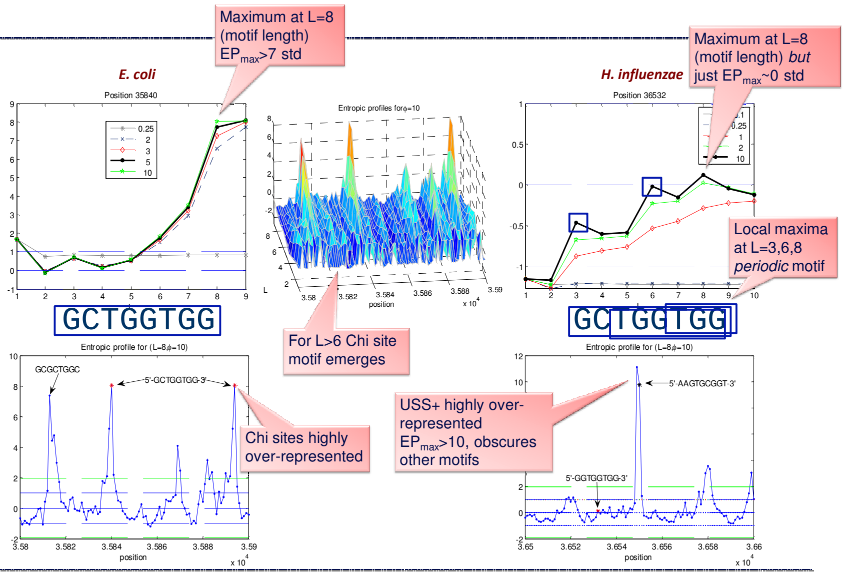
$$\boxed{\hat{f}_{L,\phi}(x_i)}=\frac{1+\frac{1}{N}\sum_{k=1}^{L}4^k\phi^k\cdot\boxed{c([i-k+1,i])}}{\sum_{k=0}^{L}\phi^k}$$

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.

4 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 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.



Maximum at L=8 (motif length) EP$_{max}$>7 std

Maximum at L=8 (motif length) *but* just EP$_{max}$~0 std

Local maxima at L=3,6,8 *periodic* motif

For L>6 Chi site motif emerges

Chi sites highly over-represented

USS+ highly over-represented EP$_{max}$>10, obscures other motifs

5 Conclusions

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

References

• Vinga, S. and Almeida, J. S. *J Theor Biol* **2004**, 231:377-388.
• Almeida, J. S. and Vinga, S. *Algorithms Mol Biol* **2006**, 1:18.
• Vinga, S. and Almeida, J.S. **2007** (submitted).

Acknowledgments

The authors acknowledge financial support by projects DynaMo (PTDC/EEA-ACR/ 69530/2006, FCT), MaGiC (IE02ID01004, INESC-ID) and the Sixth Framework Programme of the European Union - Biosapiens project.