ISMB ECCB 2009
June 27 – July 2, 2009
STOCKHOLM
U11

ENTROPIC PROFILER

Efficient whole genome analysis using information theory and statistical concepts

Francisco Fernandes[a], Ana T. Freitas[a], Jonas S. Almeida[c], Susana Vinga[a,b]

[a] *INESC-ID/IST,* Portugal [b] *FCM/UNL,* Portugal [c] *MDAnderson Cancer Center*, Houston, USA

Email: atf@inesc-id.pt and svinga@kdbio.inesc-id.pt

1 Abstract

Entropic Profiles (EP) are local information plots that indicate overall conservation of motifs in genomes. They are based on Information Theory concepts, in particular to the Renyi entropy of biological sequences. The present tool implementation, based on new data structures and algorithmic simplifications, allows to process whole genomes in few minutes.
ENTROPIC PROFILER is freely available as a web interface and downloadable source code at http://kdbio.inesc-id.pt/software/ep/ and is fully described in [1].

2 Introduction

In a recent paper [2], the authors presented the concept of Entropic Profiles (EP), a new method to extract and classify relevant and statistically significant segments of DNA sequence. The study of these motifs is very relevant because under or over-representation segments are often associated with significant biological meaning.
In this work 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 [3], using the Parzen window estimation method [4] applied to the Chaos Game Representation (CGR) of a sequence [5].

3 Methods

➢ The EP function

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

Entropic profile for the i^{th} symbol s_i, coordinate x_i
- L is the length resolution
- ϕ is a smoothing parameter

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

*can be interpreted as a linear combination of suffix counts up to a given memory length L, with increasing or decreasing weights (values of ϕ).

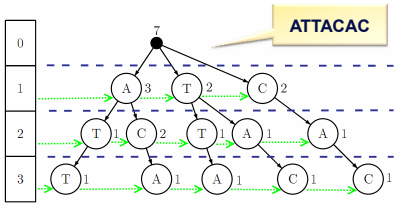
❑ To allow the comparison of different parameter sets EP values are further normalized by mean $m_{L,\phi}$ and standard deviation $s_{L,\phi}$ using all positions i=1,...,N:

$$EP_{L,\phi}(x_i) \equiv \frac{\hat{f}_{L,\phi}(x_i) - m_{L,\phi}}{s_{L,\phi}} \quad , \quad where$$

$$m_{L,\phi} = \frac{1}{N}\sum_{i=1}^{N} \hat{f}_{L,\phi}(x_i) \quad and \quad s_{L,\phi} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{f}_{L,\phi}(x_i) - m_{L,\phi}\right)^2}$$

➢ Data structure

❑ A *limited depth suffix tree* structure that stores the count of each suffix was developed to improve these computations. Algebraic simplifications further allowed efficient EP calculation



❑ The word counts are progressively updated during the building of the tree, allowing to know the exact number of occurrences of each suffix inside the whole sequence just by simple word matching operations

❑ All nodes at the same depth are connected through "side links".

➢ Performance

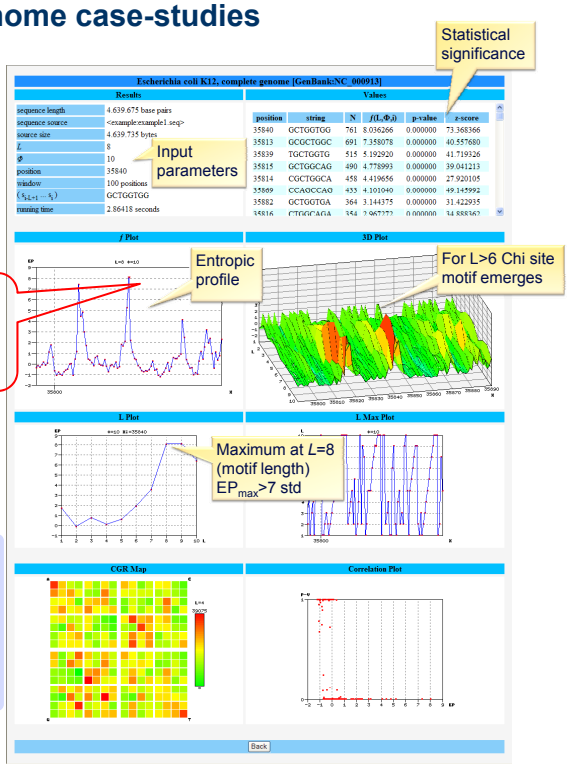
❑ Large sequences are processed very efficiently

Name	Length (Mbases)	Time (s)	Memory (MB)
Caenorhabditis elegans	97.6	75	171
Homo sapiens (CHR1)	239.2	164	290
Takifugu rubripes	382.7	263	411
Gallus gallus	940.1	611	1016
Danio rerio	1491.0	960	1567

4 Whole genome case-studies

➢ Position study

Escherichia coli genome



❑ Corresponds to a **Chi site** (Crossover Hotspot Instigator)

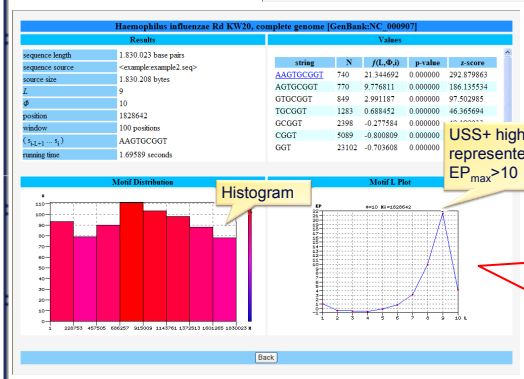
(5′-GCTGGTGG-3′)

*key region that modulates the exonuclease activity of RecBCD, an enzyme that is necessary for chromosomal dsDNA repair and integration of exogenous dsDNA)

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

➢ Motif Study

Haemophilus influenza genome



❑ Analysis of the motif which represents a **USS+** (Uptake Signal Sequence)

5′-AAGTGCCGGT-3′)

*USSs are involved in natural competence, which is a genetically controlled form of horizontal gene transfer in some bacterial species

5 Conclusions and future directions

- **Entropic profiles (EP) provide useful local information about global features of DNA. The current implementation exhibits excellent performance for sequences up to 2Gbp.**
- **Tests on whole genomes corroborate the strengths of this approach to detect biologically meaningful DNA segments, related with the detection of local scales and suffix/motifs over or under-representation**
- **Further improvements include: search for motifs in IUPAC code and analyze both forward and reverse strands and development of a web-based application with a simple Representational State Transfer (REST) I/O. As a consequence, this functionality will be also easily accessible as a URL POST which enables using the application as a web service.**

References

[1] Fernandes et al. (2009) *BMC Research Notes* 2:72.
[2] Vinga S, Almeida JS (2007) *BMC Bioinformatics* 8:393.
[3] Rényi A (1961) *University of California Press* 547-561.
[4] Parzen E (1962) *The Annals of Math Stat* 33:1065-1076.
[5] Jeffrey HJ (1990) *Nucleic Acids Res* 18:2163-2170.

Acknowledgments

The authors acknowledge financial support by projects DynaMo (PTDC/EEA-ACR/ 69530/2006, FCT) and ARN (PTDC/EIA/67722/2006, FCT).