BMC Research Notes

BioMed Central

Technical Note

Entropic Profiler – detection of conservation in genomes using information theory

Francisco Fernandes[1], Ana T Freitas[1,2], Jonas S Almeida[3] and Susana Vinga*[1,4]

Address: [1]Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento (INESC-ID), R Alves Redol 9, 1000-029 Lisboa, Portugal, [2]Instituto Superior Técnico – Universidade Técnica de Lisboa (IST/UTL), Av Rovisco Pais, 1049-001 Lisboa, Portugal, [3]Dept Biostat Appl Math, Univ Texas MDAnderson Cancer Center – unit 447, 1515 Holcombe Blvd, Houston TX 77030-4009, USA and [4]Dept Bioestatística e Informática, Faculdade de Ciências Médicas – Universidade Nova de Lisboa (FCM/UNL), Campo Mártires da Pátria 130, 1169-056 Lisboa, Portugal

Email: Francisco Fernandes - francisco.fernandes@ist.utl.pt; Ana T Freitas - atf@kdbio.inesc-id.pt; Jonas S Almeida - jalmeida@mdanderson.org; Susana Vinga* - svinga@kdbio.inesc-id.pt

* Corresponding author

Published: 5 May 2009

BMC Research Notes 2009, **2**:72 doi:10.1186/1756-0500-2-72

Received: 4 December 2008
Accepted: 5 May 2009

This article is available from: http://www.biomedcentral.com/1756-0500/2/72

Abstract

Background: In the last decades, with the successive availability of whole genome sequences, many research efforts have been made to mathematically model DNA. Entropic Profiles (EP) were proposed recently as a new measure of continuous entropy of genome sequences. EP represent local information plots related to DNA randomness and are based on information theory and statistical concepts. They express the weighed relative abundance of motifs for each position in genomes. Their study is very relevant because under or over-representation segments are often associated with significant biological meaning.

Findings: The Entropic Profiler application here presented is a new tool designed to detect and extract under and over-represented DNA segments in genomes by using EP. It allows its computation in a very efficient way by recurring to improved algorithms and data structures, which include modified suffix trees. Available through a web interface http://kdbio.inesc-id.pt/software/ep/ and as downloadable source code, it allows to study positions and to search for motifs inside the whole sequence or within a specified range. DNA sequences can be entered from different sources, including FASTA files, pre-loaded examples or resuming a previously saved work. Besides the EP value plots, p-values and z-scores for each motif are also computed, along with the Chaos Game Representation of the sequence.

Conclusion: EP are directly related with the statistical significance of motifs and can be considered as a new method to extract and classify significant regions in genomes and estimate local scales in DNA. The present implementation establishes an efficient and useful tool for whole genome analysis.

Findings

Algorithm

In a recent paper [1], 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.

The concept of Entropic Profiles was proposed previously but in a different context and distinct scope [2]. In that work, EP were calculated using the Shannon entropies of the histograms obtained from the Chaos Game Representation (CGR) [3] for different resolutions, i.e., using different L-tuples. This procedure allowed the discrimination between random and natural DNA sequences. Although the same name was used, this previous endeavor focused on a global perspective of sequence entropy [4] whereas the present work investigates a local entropy formulation. Another type of sequence profile also previously explored was based on Sequence Logos [5], giving the information content per position. However, this procedure required the alignment of a set of sequences and therefore is not an individual-based, alignment-free approach [6,7] like the present one.

Another type of sequence profile also explored was based on linguistic complexity [8] and low entropy DNA zones [9]. More recently, compression schemes were used to define Information Sequences (IS) [10] and the linear space complexity obtained allows exploring DNA. This algorithm uses a Lempel-Ziv-like compression algorithm to keep track of the repetitions, which can be very efficiently implemented by using suffix trees [11], also present in Entropic Profiler. However, this information content approach does not take into account neither explores the fact that a repetition can occur at different depths/lengths simultaneously since it uses a fixed low-order Markov model, geometric distribution and global set of probabilities for the beginning and ending of the repetitions. Entropic profiles, on the other hand, can automatically detect the best length to use in each position and also allow the study of a specific resolution while providing at the same time an overview of a range of lengths through a 3D plot. In this sense EP represent a combinatorial and exact approach, whereas IS are based in given compression schemes and Bayesian statistical models.

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 [12], using the Parzen window estimation method [13] applied to the CGR of a sequence [3]. 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.

To our knowledge, there are no current methods that calculate local information plots looking just at one individual sequence. The local scale given by the EP are estimated from the complete unique sequence and therefore can putatively have unbiased information about particular positions in genomes, without using any additional external information. The closest method is related to the calculation of the statistical significance of motifs, given as p-values or z-scores [14].

The main EP function [1] is defined by:

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

where L is the length resolution chosen, φ is a smoothing parameter, and $c([i\text{-}k+1,\ i])$ is the number of motifs ($x_{i\text{-}k+1}$ ... x_i) in the whole sequence x_1 ... x_N, i.e., the number of occurrences of the substring of length k that ends at position i in the sequence. This function can be interpreted as a linear combination of suffix counts up to a given memory length L, with increasing and decreasing weights (values of φ). It retrieves, for each position in the sequence, information about the abundance of the corresponding L-tuple suffix inside the entire sequence. To allow the comparison of different parameter combinations EP values must be further normalized by mean $m_{L,\ \varphi}$ and standard deviation $s_{L,\ \varphi}$ using all positions i = 1,..., N:

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

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

The computation (Eq. 1 and 2) is very time consuming because obtaining $m_{L,\ \varphi}$ and $s_{L,\ \varphi}$ involves calculating the EP values for all the sequence positions and all the calculations have to be executed several times and repeated for each value of L considered.

Along with this new method [1], the authors supplied a computational application developed in MATLAB to demonstrate their approach. Although the command-line based executables were also provided, the prototype version of this application provided a very poor performance, with excessive running times and memory usage. These restrictions severely hamper the usability of entropic profiles in practical situations, e.g., whole genome analysis. Furthermore, no additional information was provided regarding the comparison with other methods.

In the present implementation all the formulas were restructured and subject to several optimizations, which included the use of new data structures and extensive algebraic simplifications (see Supplementary Material). The result was a highly efficient web-based application, now described.

Implementation

• Data structures

The major computational challenge in the core EP function corresponds to operations over suffixes with limited lengths. To improve these computations an efficient limited depth suffix tree structure [15] that stores the counts of each suffix was developed (Figure 1). The depth limit prevents the tree from growing without bounds, reducing time and memory consumption. The word counts are progressively updated during the building of the tree, allowing knowing the exact number of occurrences of a particular suffix inside the whole sequence just by a simple word matching operation over the suffix tree.

Furthermore all nodes, in the tree, at the same depth are connected through "side links" which are very useful to speed up some calculations.

This procedure is similar to creating a k-truncated suffix tree [16-18], which is demonstrated to efficiently searching for specific strings of length of up to *k* characters in large sequences. The present proposal has the difference that each node is labeled with a single symbol, so it resembles more to a trie than to a tree. Although this makes its construction a little slower and less compact, the speed improvement of this structure is noticed when we need the counts of all substrings of the same length, efficiently retrieved using the "depth-links".

• Algebraic and algorithm simplifications

The major improvement in all the simplifications on the formulas in Eq. 1 and Eq. 2 is that one can scan through all the motifs of a particular length inside the suffix tree instead of scanning through all the positions inside the whole sequence. The number of distinct motifs of length *L* is definitely much smaller than the size of the sequence. The suffix tree allows the indexation of all the L-tuple words and all their corresponding suffixes, for easy retrieval, and stores all their counts by scanning through the sequence only once. On the other hand, the "side links" allow to traverse the suffix tree horizontally and cover the counts of all words of the same length in a single sweep instead of searching the whole sequence over and over again.

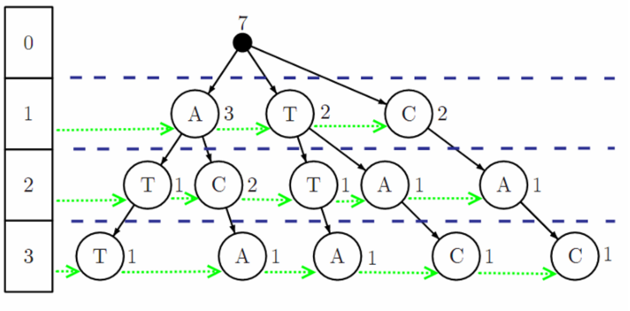


Figure 1

Suffix tree and side links. Example of a suffix tree and side links for word "ATTACAC" showing the suffix counts limited to depth 3. All substrings of length 3 are represented in the tree ("ATT", "TTA", "TAC", "ACA", "CAC"). Also shown are the side links connecting nodes of the same depth.

Because DNA sequences can be extremely large, the search procedure for the positions of a particular motif has to be very efficient: the fast exact string matching algorithm "Shift-And" [19] was used for this purpose. Its great speed comes from the fact that all the operations required are implemented using basic arithmetic bitwise operations. All of these algorithms and data structures combined offer a fast, solid and robust backbone for the Entropic Profiler application.

• Statistical significance of motifs

In order to allow the comparison of the EP values with other statistical efforts, p-values and z-scores of the analyzed motifs are also reported. Over-represented motifs have a very low p-value and very high z-scores and EP values. These values are calculated using first-order Markov chain transition probability tables estimated directly from the whole sequence taking into account the overlap capacity or period of each motif [14].

• Interface

The Entropic Profiler tool is presented through an intuitive and easy to use web interface. The user can input a sequence by copy/paste the text or by uploading a FASTA file format. The sequence is saved so the user can later return to previous work. Pre-loaded examples are also available. The study can then be performed by selecting a particular position or a specific motif.

By default the results are displayed to the web browser. If the job takes more than 30 seconds the user is informed that an email will be sent with a link to the results, which will be kept by the system for five days. The Entropic Profiler tool provides different plots that allow the user to visually inspect information about all the motifs of distinct lengths within a given range, in a way that the most relevant ones are easily detected. EP values and detailed statistical data are also provided for a more precise reference. All the information can be downloaded as text files for subsequent study.

Testing

The Entropic Profiler was tested in several DNA sequences, and the results for two genomes will be described below as two distinct examples of application. In particular, the user input options will be presented and the corresponding output will be described. After each example the user can click on the "back" button and recover the previous job for further analysis.

1) Position study – Escherichia coli genome

When studying a position, the motif length can be chosen or the application can automatically determine the best one to use. A window length can also be specified to study a determined range of bp around the position.

Figure 2 shows the output results for a pre-loaded example of the *Escherichia coli K12* genome which uses as (default) input parameters the values $L = 8$, $\varphi = 10$, position 35840 and a window length of 100. Panel A) has the input values given from the user, for reference. In panel B) the listing of the all the motifs (string) with EP values $f(L, \varphi, i)$ are given in decreasing order, along with their positions in the sequence and the corresponding p-values and z-scores. The user can download all the table information by clicking in the corresponding button, repeat the analysis for other position (click on the desired one) or perform a study by motif (see next example). Panel C) represents the EP plot along the range positions previously defined, where higher values correspond to the most significant suffixes that end in that position. Some information appears by pointing the mouse to each figure. Panel D) is the 3D plot of the same information but allowing for the variation of parameter L, the resolution. From this, one can analyze the relative increase of EP and visually identify the most promising L values that will lead to longer or more significant motifs. Panel E) plots the EP values for one particular position as a function of L, showing that there are values that maximize (or minimize) the EP, i.e., which correspond precisely to a local scale or the most distinctive suffix, when compared with the overall sequence. This allows to chose the "best" L for that particular position, which can be analyzed further. Panel F) calculates the value L, for each position in the defined range, which maximizes EP, which represents the local scale of each position. Panel G) is the CGR map of the complete sequence for resolution 4, which shows the relative density of strings. Finally, panel H) is a scatter plot that represents the relation between EP and p-values for the collection of motifs in the specified range.

A further look at this output shows that at position 35840 a highly rated motif of length 8 can be found (*5'-GCTGGTGG-3'*), which corresponds to a Chi site (crossover hotspot instigator), a key region that modulates the exonuclease activity of RecBCD (an enzyme that is necessary for chromosomal dsDNA repair and integration of exogenous dsDNA) [20-22].

It is noteworthy that this position can be obtained by analyzing the values L that maximize EP, without a previous knowledge of the length of the motif under study. In fact, each individual position corresponds to a local scale given by the resolution L that maximizes EP in that position, which equals exactly $L = 8$, the length of the biologically relevant motif.

2) Motif Study – Haemophilus influenza genome

When studying the distribution of a specific motif, the application finds all the occurrences of the motif in the whole sequence or within a defined search range and plots the corresponding histogram in the submitted strand.

Figure 3 shows the analysis for the well known motif "AAGTGCGGT" [23] in *Haemophilus influenza*, which represents a uptake signal sequence (USS+). USSs are involved in natural competence, which is a genetically controlled form of horizontal gene transfer in some bacterial species, related to their ability to take up DNA from the surrounding environment (reviewed in [24]).

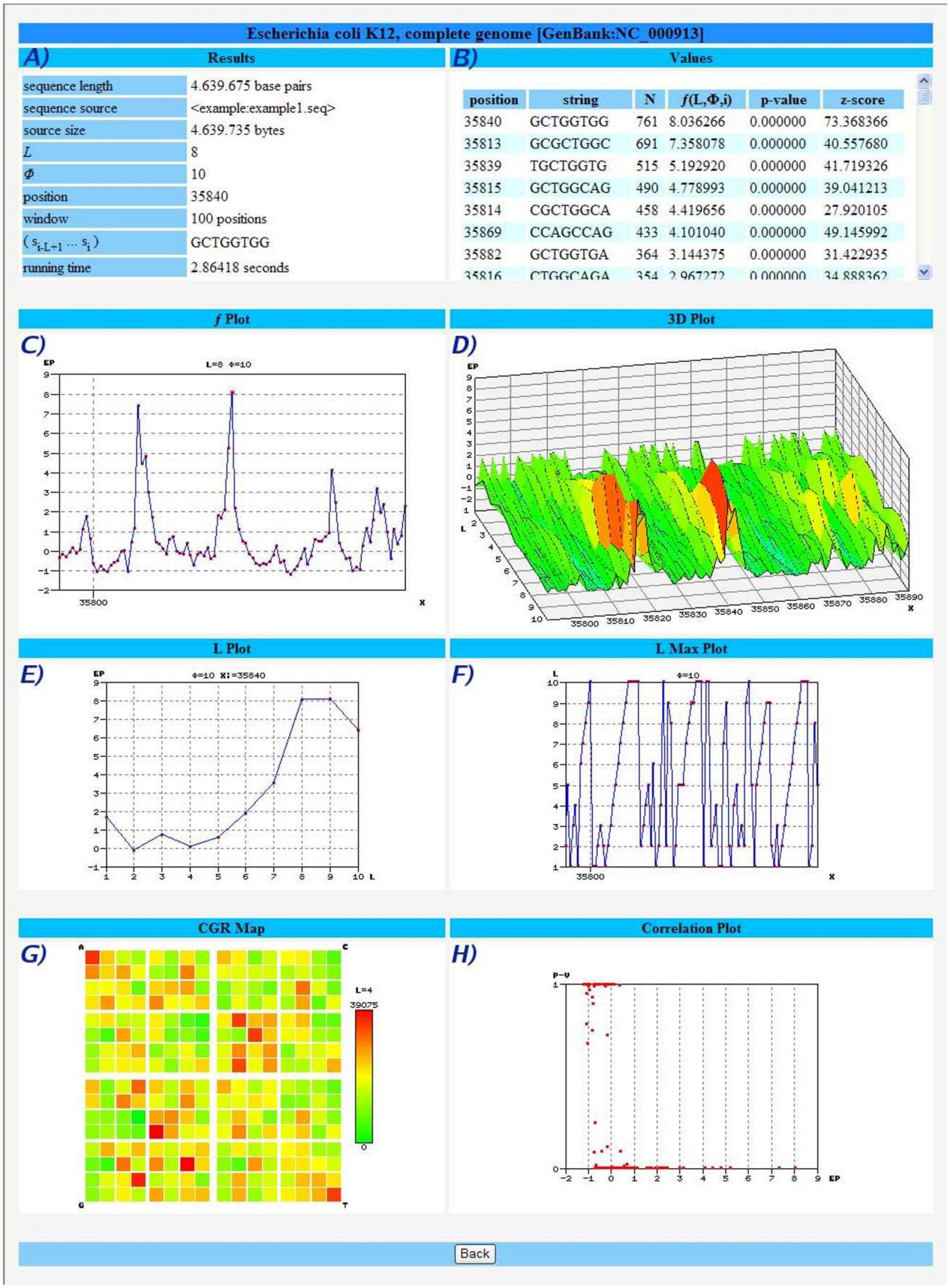
Panel A) tables the input values given by the user. Panel B) represents the information for the input motif, along with all its sub-motifs down to length 3 (GGT in this case); the EP, p-values and z-score are calculated and the user can further click on each string to get the list of all positions where that string appears in the sequence (first symbol). All the information is available for download. These positions can be further studied by using the procedure above described in Example 1. The histogram of the occurrences of the motif is represented in panel C) and can be used to analyze the distribution of the positions of occurrences, shown to be an important feature of biologically significant motifs [25]. Finally, panel D) analyses one of the positions of occurrence in terms of the EP values vs. parameter L, which allows the determination of important resolutions for that particular string.

If the length of the motif is not known in advance, the user can search for particular maxima along the sequence – these positions can be further analyzed and the corresponding L retrieved, after which a more meticulous analysis can be conducted for those local scales.

• *Performance*

The average processing speed is around 2 seconds per million of basepairs (running on a machine with an Intel Xeon Quad-core CPU @ 1.60 GHz with 8 GB RAM). When loading a previous job or example these times are greatly decreased since the necessary data structures are already built and stored.

The current web application allows a maximum file size of 5 MB due to PHP and webserver constraints. For longer sequences a local version of Entropic Profiler should be used, by downloading the source code available at the webpage.



Escherichia coli K12, complete genome [GenBank:NC_000913]

Results

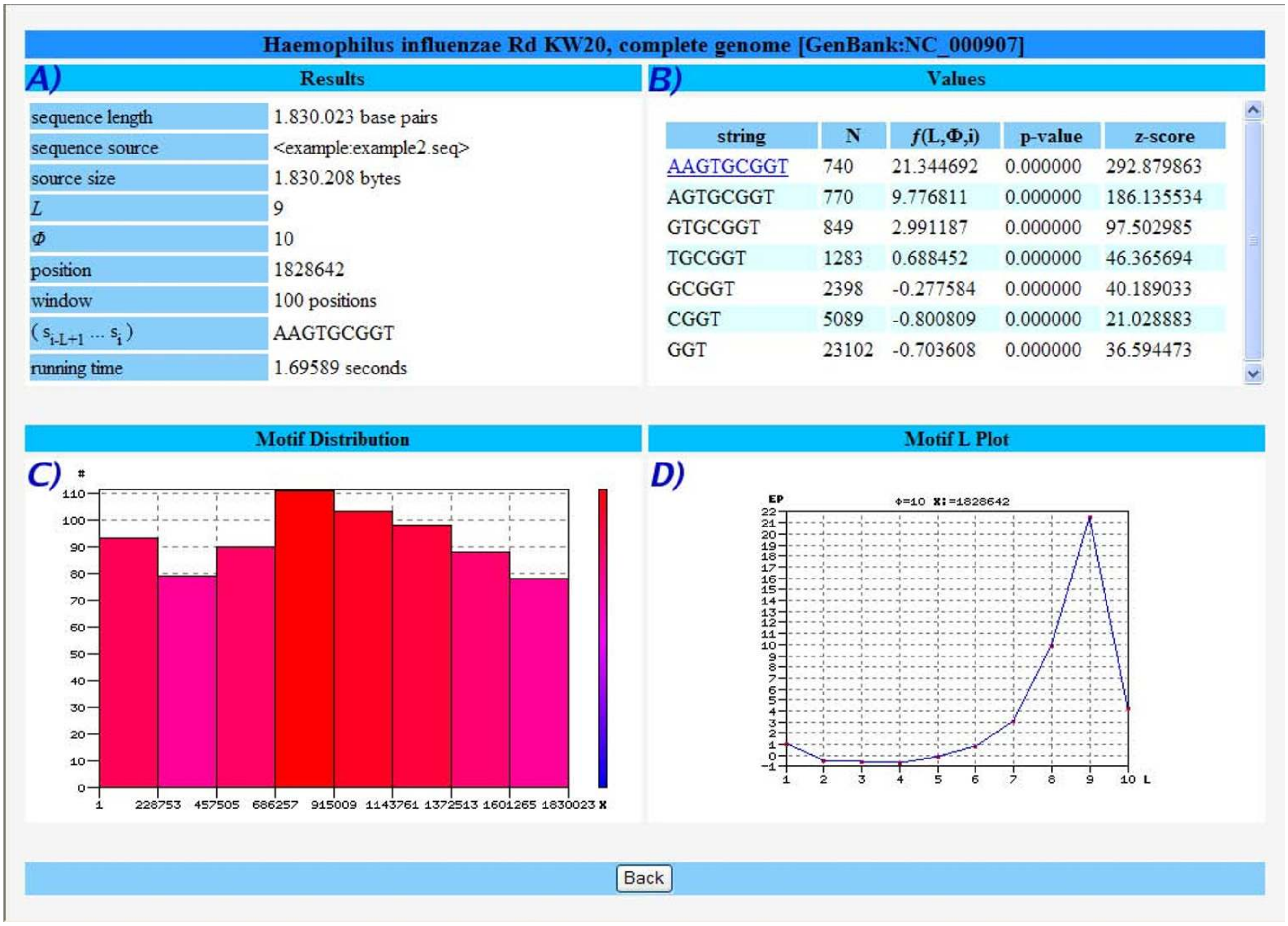
sequence length	4.639.675 base pairs
sequence source	<example:example1.seq>
source size	4.639.735 bytes
L	8
Φ	10
position	35840
window	100 positions
($s_{i-L+1} \cdots s_i$)	GCTGGTGG
running time	2.86418 seconds

Values

position	string	N	$f(L,\Phi,i)$	p-value	z-score
35840	GCTGGTGG	761	8.036266	0.000000	73.368366
35813	GCGCTGGC	691	7.358078	0.000000	40.557680
35839	TGCTGGTG	515	5.192920	0.000000	41.719326
35815	GCTGGCAG	490	4.778993	0.000000	39.041213
35814	CGCTGGCA	458	4.419656	0.000000	27.920105
35869	CCAGCCAG	433	4.101040	0.000000	49.145992
35882	GCTGGTGA	364	3.144375	0.000000	31.422935
35816	CTGGCAGA	354	2.967272	0.000000	34.888362

Figure 2

Study sequence by position. Example of the input and output screens of the Entropic Profiler application while studying around a position in the *Escherichia coli* K12 genome.



Haemophilus influenzae Rd KW20, complete genome [GenBank:NC_000907]

Results

sequence length	1.830.023 base pairs
sequence source	<example:example2.seq>
source size	1.830.208 bytes
L	9
Φ	10
position	1828642
window	100 positions
($s_{i-L+1} \cdots s_i$)	AAGTGCGGT
running time	1.69589 seconds

Values

string	N	f(L,Φ,i)	p-value	z-score
AAGTGCGGT	740	21.344692	0.000000	292.879863
AGTGCGGT	770	9.776811	0.000000	186.135534
GTGCGGT	849	2.991187	0.000000	97.502985
TGCGGT	1283	0.688452	0.000000	46.365694
GCGGT	2398	-0.277584	0.000000	40.189033
CGGT	5089	-0.800809	0.000000	21.028883
GGT	23102	-0.703608	0.000000	36.594473

Figure 3
Study sequence by motif. Example of the input and output screens of the Entropic Profiler application while searching for motif "AAGTGCGGT" on *H. influenzae*.

The examples provided above are for short bacterial genomes. This was in order to test Entropic Profiler in the analysis of known biologically relevant motifs. However, the analysis of large genomes can pose several problems, such as memory usage. In this context, compressed suffix arrays constructed with Burrows-Wheeler Transform were previously used to create indexes which allow efficient analysis [26,27]. Since Entropic Profiler uses suffixes trees which are limited in depth, a memory-wise efficient solution is also obtained, as described above.

In order to assess the performance of the tool for larger genomes, we also ran the algorithm for several other species several orders longer. The results, in terms of speed and memory, are in Table 1.

As the results show, the additional space used beyond the size of the sequence is almost unnoticeable. This is explained by the fact that the suffix tree structure used has a limited depth, therefore it does not grow exponentially.

Although no biological analysis was performed in this situation, which would be out of the scope of this manuscript, these results demonstrate the feasibility of this type of analysis for large genomes and validate Entropic Profiler application in more challenging data.

Future directions

Possible improvements to be implemented include the possibility to search for motifs in IUPAC code and the option to analyze both the forward and reverse strand of a sequence. The choice of programming language and web-based deployment was geared towards the future development of a web-based application with a simple Representational State Transfer (REST) I/O. As a consequence, the functionality described in this report will be

Table 1: Performance of Entropic Profiler for larger genomes.

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

For each species name, the length of each sequence tested (in 10^6 bases), the total running time (in seconds) and the memory usage (megabytes) are reported.
Note: The sequences can be assessed in the following links: Caenorhabditis elegans ftp://ftp.wormbase.org/pub/wormbase/genomes/c_elegans/sequences/dna/c_elegans.WS199.dna.fa.gz; Homo sapiens (CHR1) ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/CHR_01/hs_ref_chr1.fa.gz; Takifugu rubripes ftp://ftp.jgi-psf.org/pub/JGI_data/Fugu/v4.0/Takru4.allmasked.gz; Gallus gallus ftp://ftp.ncbi.nlm.nih.gov/genomes/Gallus_gallus/pre-release/contig_fasta/ref_placed.fa.gz; Danio rerio ftp://ftp.ensembl.org/pub/assembly/zebrafish/Zv8release/Zv8_contigs.fa.

also easily accessible as a URL POST which enables using the application as a web service.

Conclusion

This work focused on the development of an efficient tool available through a web interface to compute the entropic profiles values for given DNA sequences. Since the main expression used to compute them is a function of suffix counts, the use of *suffix trees* to represent the sequences turned to be the major performance boost over the previous application.

Thanks to the developed side-links structure, calculations that involved searching the whole sequence over and over again to retrieve the counts of words of a particular length are now replaced by a single horizontal scan of the suffix tree at the depth corresponding to that specific word length. Some of the accomplished ideas, algorithms or structures developed for this particular work can also be reused and later readapted for building future efficient applications in similar sequence analysis projects. Besides the EP value plots, the p-value and z-score for each motif are also computed, along with the Chaos Game Representation of the sequence. The ability to load the sequence from different sources and to restore the last work also improves greatly the usability of the application. Its structured web-based presentation gave it a more appealing look and its intuitive interface made its use extremely straightforward. All of this makes the described application a fast, efficient and extremely useful tool for sequence and motif analysis that can be further expanded and combined with other algorithms.

Availability and requirements

Project name: Entropic Profiler

Project home page: http://kdbio.inesc-id.pt/software/ep/

Operating system(s): Platform independent

Programming language: C

Other requirements: None

License: None

Any restrictions to use by non-academics: None

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FF designed and implemented the algorithm optimizations and data structures and coded the application and the web interface. ATF and SV introduced the idea of applying suffix trees to EPs and provided insights and suggestions to improve the application. SV and JSA introduced the EP theory and guided the application development process. All authors contributed for the writing, read and approved the final manuscript.

Acknowledgements

We thank INESC-ID for providing computing resources. This project was supported by projects ARN – Algorithms for the identification of genetic Regulatory Networks (PTDC/EIA/67722/2006) and DynaMo – Dynamical modeling, control and optimization of metabolic networks (PTDC/EEA-ACR/69530/2006) from the Portuguese Science Foundation (FCT). The authors would like to thank the referees whose useful suggestions greatly improved this manuscript.

References

1. Vinga S, Almeida JS: **Local Rényi entropic profiles of DNA sequences.** *BMC Bioinformatics* 2007, **8:**393.
2. Oliver JL, Bernaola-Galvan P, Guerrero-Garcia J, Roman-Roldan R: **Entropic profiles of DNA sequences through chaos-game-derived images.** *J Theor Biol* 1993, **160:**457-470.
3. Jeffrey HJ: **Chaos Game Representation of Gene Structure.** *Nucleic Acids Res* 1990, **18:**2163-2170.
4. Vinga S, Almeida JS: **Rényi continuous entropy of DNA sequences.** *J Theor Biol* 2004, **231:**377-388.
5. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: **Information content of binding sites on nucleotide sequences.** *Journal of molecular biology* 1986, **188:**415-431.
6. Vinga S: **Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for DNA and protein classification.** In *Advanced Computational Methods for Biocomputing and Bioimaging* Edited by: Pham TD, Yan H, Crane DI. New York: Nova Science Publishers; 2006.
7. Vinga S, Almeida J: **Alignment-free sequence comparison – a review.** *Bioinformatics* 2003, **19:**513-523.

8. Troyanskaya OG, Arbell O, Koren Y, Landau GM, Bolshoy A: **Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity.** *Bioinformatics* 2002, **18:**679-688.
9. Crochemore M, Verin R: **Zones of low entropy in genomic sequences.** *Comput Chem* 1999, **23:**275-282.
10. Dix TI, Powell DR, Allison L, Bernal J, Jaeger S, Stern L: **Comparative analysis of long DNA sequences by per element information content using different contexts.** *BMC Bioinformatics* 2007, **8(Suppl 2):**S10.
11. Larsson NJ: **Extended application of suffix trees to data compression.** *DCC '96 – Data Compression Conference, Proceedings* 1996:190-199.
12. Rényi A: **On measures of entropy and information.** In *Proc of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability; June 20 – July 30, 1960* University of California Press; 1961:547-561.
13. Parzen E: **On Estimation of a Probability Density Function and Mode.** *The Annals of Mathematical Statistics* 1962, **33:**1065-1076.
14. Robin S, Rodolphe F, Schbath S: *DNA, words, and models* New York, NY: Cambridge University Press; 2005.
15. Gusfield D: *Algorithms on strings, trees, and sequences: computer science and computational biology* Cambridge [England]; New York: Cambridge University Press; 1997.
16. Schulz MH, Bauer S, Robinson PN: **The generalised k-Truncated Suffix Tree for time-and space-efficient searches in multiple DNA or protein sequences.** *International journal of bioinformatics research and applications* 2008, **4:**81-95.
17. Allali J, Sagot MF: **The at most k-deep factor tree.** *Technical Report Institut Gaspard-Monge, Université de Marne la Vallée, IGM 2004–03* 2004.
18. Na JC, Apostolico A, Iliopoulos CS, Park K: **Truncated suffix trees and their application to data compression.** *Theoretical Computer Science* 2003, **304:**87-101.
19. Wu S, Manber U: **Fast Text Searching Allowing Errors.** *Communications of the ACM* 1992, **35:**83-91.
20. Arakawa K, Uno R, Nakayama Y, Tomita M: **Validating the significance of genomic properties of Chi sites from the distribution of all octamers in Escherichia coli.** *Gene* 2007, **392:**239-246.
21. Sourice S, Biaudet V, El Karoui M, Ehrlich SD, Gruss A: **Identification of the Chi site of Haemophilus influenzae as several sequences related to the Escherichia coli Chi site.** *Mol Microbiol* 1998, **27:**1021-1029.
22. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al.*: **The complete genome sequence of Escherichia coli K-12.** *Science* 1997, **277:**1453-1474.
23. Karlin S, Mrazek J, Campbell AM: **Frequent oligonucleotides and peptides of the Haemophilus influenzae genome.** *Nucleic Acids Res* 1996, **24:**4263-4272.
24. Dubnau D: **DNA uptake in bacteria.** *Annu Rev Microbiol* 1999, **53:**217-244.
25. Casimiro AC, Vinga S, Freitas AT, Oliveira AL: **An analysis of the positional distribution of DNA motifs in promoter regions and its biological relevance.** *BMC Bioinformatics* 2008, **9:**89.
26. Lippert RA: **Space-efficient whole genome comparisons with Burrows-Wheeler transforms.** *J Comput Biol* 2005, **12:**407-415.
27. Healy J, Thomas EE, Schwartz JT, Wigler M: **Annotating large genomes with exact word matches.** *Genome research* 2003, **13:**2306-2315.