Session:

Proceedings of the 6th conference of the
Hellenic Society for Computational Biology & Bioinformatics - HSCBB11
University of Patras Conference Center, 7-9/10/2011

XX/10/2011

Analyzing the DNA composition of ultraconserved sequences with N-gram Graphs.

D. Polychronopoulos[1], G. Giannakopoulos[2,*], C. Nikolaou[1,3,*], G. Paliouras[2] and Y. Almirantis[1]

[1]Institute of Biology, NCSR "Demokritos"

[2]Institute of Informatics and Telecommunications, NCSR "Demokritos"

[3]Department of Biology, University of Crete

[*]Correspondence to: cnikol@bio.demokritos.gr, ggianna@iit.demokritos.gr

ABSTRACT

Motivation: One of the most striking discoveries to have emerged from comparisons among mammalian and other genomes is the existence of hundreds of noncoding elements of more than 200 bp in length that show absolute identity among mammalian orders [1]. These elements represent the tip of the iceberg of a much larger class of conserved noncoding elements (CNEs). There have been many speculations about what the exact role of these elements may be, proposing that they might act as enhancers [2] or even insulators [3]. The diverse nature of these elements appears to be in contrast with their very particular DNA composition, the study of which may provide us with insight on their possible functional roles, their genomic distribution and evolutionary background.

Methods: Ultraconserved sequences for *H.sapiens* and *C. elegans* were obtained and analyzed through the N-gram Graph approach [4]. The N-gram graphs (NGG) represent how symbols (e.g., nucleotides) co-occur within a given neighborhood (e.g., within an oligonucleotide). The neighborhood is defined based on a distance function (e.g., a neighborhood of 5 consecutive characters within a text). Under this framework we trained graphs with the UCS and compared them with genomic and random surrogate sequences with similar DNA composition, in order to define specific "rules" in the use of nucleotides existing within UCS.

Results: NGG-assisted classification of UCS and CNEs (sequences obtained with less stringent conservation criteria) against random genomic sequences was accurate at a rate of 76%. CNEs were better distinguished from random sequences with identical composition than from natural genomic sequences with the same GC content. Furthermore, a classification of different collections of conserved non-coding sequences from *H. sapiens* and *C. elegans* revealed differences among sequences at two levels. On one hand, an expected discrimination between the two species was observed, on the other, ultraconserved human sequences showed marked differences in the nucleotide composition when compared to less conserved ones from the same organism. Classification of the three classes yielded an average AUC of 0.94, which greatly exceeds the one obtained with the use of conventional schemes for base composition comparisons.

Discussion: Our results suggest that the N-gram Graphs represent a very promising approach towards the study of nucleotide composition in relation to the genomic sequence functionality. Sequence classification enabled us not only to distinguish conserved non-coding sequences from random genomic fragments but in addition to observe differences within them, an observation that may well reflect diverse functional roles for these sequences. The presented analysis is being extended in order to allow for the detection of similar functional elements in raw genomic sequences.

REFERENCES

1. Bejerano, G., et al., *Ultraconserved elements in the human genome.* Science, 2004. **304**(5675): p. 1321-5.
2. Visel, A., et al., *Ultraconservation identifies a small subset of extremely constrained developmental enhancers.* Nat Genet, 2008. **40**(2): p. 158-60.
3. Xie, X., et al., *Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.* Proc Natl Acad Sci U S A, 2007. **104**(17): p. 7145-50.
4. Giannakopoulos G., K.V., Vouros G. and Stamatopoulos, P. , *Summarization system evaluation revisited: N-gram graphs.* ACM Trans. Speech Lang. Process., 2008. **5**(3): p. 1-39.