COMPARATIVE MOTIF DISCOVERY IN THE CLOUD

Dieter De Witte[*,1], *Jan Van de Velde*[2,3], *Michiel Van Bel*[2,3], *Pieter Audenaert*[1], *Piet Demeester*[1], *Bart Dhoedt*[1], *Klaas Vandepoele*[2,3] *and Jan Fostier*[1]
Dept. of Information Technology (INTEC), Ghent University[1], *Dept. of Plant Systems Biology, VIB, Ghent*[2], *Dept. of Plant Biotechnology and Bioinformatics, Ghent University*[3]
[*]*dieter.dewitte@intec.ugent.be*

We present a novel method for the computational discovery of cis-regulatory elements ('motifs') in genomic sequences based on phylogenetic footprinting. Word-based, exhaustive approaches are among the best performing methods; however, they pose significant computational challenges as the number of candidate motifs to evaluate is very high. We describe a parallel, distributed-memory algorithm for the *de novo* comparative motif discovery that has been implemented in Hadoop's MapReduce framework in order to make efficient use of cloud computing infrastructure. It is used in a comparative study of four Monocot plant species and is able to statistically evaluate the conservation of billions of candidate motifs in less than 34 hours using 20 node instances on the Amazon EC2.

INTRODUCTION

Many computational methods exist for the discovery of cis-regulatory elements, see e.g. [1] for an overview. As sequence information is available for an increasing number of organisms, methods based on phylogenetic footprinting are becoming increasingly attractive. In this contribution, we present such method, called BLSSpeller, with three important and unique features. First, it relies on a word-based methodology in which *all* words that occur in any of the sequences are exhaustively screened for conservation. In contrast to statistical methods, the method yields complete and optimal results. Second, the algorithm does not rely on pre-generated multiple sequence alignments (MSA). This ensures the method is suitable even for diverged species for which the generation of a MSA is difficult. Third, in order to deal with the very high runtimes and memory requirements associated with exhaustive, word-based methodologies, the algorithm is implemented in the MapReduce framework, in order to take advantage of cloud infrastructure such as the Amazon EC2.

METHODS

Orthologous and paralogous genes from related species are grouped into gene families. The promoter sequences are extracted and serve as input for the algorithm. For each gene family individually, all words that are conserved within that family are exhaustively enumerated using Sagot's algorithm [2]. Conservation is scored using the Branch Length Score (BLS) [3] which takes the relative evolutionary distance between the organisms into account. Words are enumerated in the IUPAC alphabet with the exclusion of three-fold degenerate characters (B, D, H, V). For each word, the number of gene families in which the word is conserved with a BLS higher than a pre-specified threshold is counted. This number is compared to a background model, i.e., the expected number of gene families in which permutations of that word (i.e., same length, base pair composition and degeneracy) is conserved. For each word, a confidence score C is established, only motifs with a confidence $C > 0.9$ are retained [3].

In order to deal with the very high number of words, the algorithm was implemented using the MapReduce framework. During the map phase, different gene families are attributed to different map tasks and each map task enumerates all conserved words. As the number of words produced by a map task is typically too high to store in memory, the words are written to local disc. In the reduce phase, the conservation score C is established for each word. In between the map and reduce phase, words are sorted on disc in a distributed fashion.

RESULTS & DISCUSSION

The input consists of four related Monocot species: *Zea mays*, *Sorghum bicolor*, *Brachypodium distachyon* and *Oryza sativa*. Using the integrative orthology method of Plaza 2.5 [4], 17724 gene families were constructed. For each gene, the 2kbp promoter was extracted. The software was run on the Amazon EC2 cloud on 20 nodes (type m1.xlarge). The total runtime was 33 hours and 28 minutes. In total, over 2.4 trillion candidate motifs with a length between 6 to 12 basepairs and a maximum of three degenerate characters were emitted by the mappers. This corresponds to 3.65 TByte of data that was sorted on disc in between the map and reduce phase. Over 620 million words with a confidence score $C > 0.9$ were retained by the reduce tasks. This corresponds to roughly 4% of the total number of unique words examined. The identified motif instances show a significant enrichment towards (experimentally characterized) open chromatin regions in rice seedling (12,59 fold enrichment, $p < 0.001$), revealing their biological relevance. Additionally, the method correctly predicts a set of experimentally determined Knotted1 gene targets that were obtained using ChIP-Seq combined with transcript profiling in *Zea mays* [5]. Future work consists of the development of post-processing tools by which the output of the discovery algorithm can be queried in order to identify specific regulatory interactions.

REFERENCES

1. Das M.K. & Dai H.-K. *BMC bioinformatics* **8** Suppl. 7, S21 (2007).
2. Marsan L. and Sagot M. F. Journal of computational biology 7, 325-362 (2000).
3. Kheradpour P. *et al. Genome research* **17**(12), 1919-1931 (2007).
4. Van Bel *et al. Plant Physiology*, 111.189514 (2011).
5. Bolduc N. *et al. Genes Dev.* **26**(15), 1685-1690, (2012).