Title

Matthias Grenié[1], Jean-François Goût[2], Michael Lynch[2]
1 Départment de Biologie, École Normale Supérieure de Lyon, Lyon, France
2 Biology Department, Indiana University, IN, United States of America

Abstract

Author Summary

Introduction

Structure of the introduction

- Whole Genome Duplications background, major evolutionary force
- the *Paramecium* project, why *Paramecium* is interesting, the aurelia complex
- Here, focusing on the computational part, developing pipeline, showed that etc.

Since Ohno first hypothesized the influence of Whole-Genome Duplications (WGD), scientists kept showing that a broad number of families experienced at least one WGD : yeast, insects, Angiosperma, Vertebrates, Salmonids, and many others. WGD are evolutionary event when the genome of a given individual is duplicated, meaning that the whole genome is in two copies, duplicated pairs of genes are **paralogs**. WGD may be involved in many evolutive radiations as it creates a context of loosen selection. According to the Duplication-Degeneration-Complementation model,

To understand the consequences of WGDs we have been studying the *Paramecium aurelia* complex. *Paramecium* are a Ciliates group. (See position on phylogenetic tree) As one of the only free-living eukaryotes studied, other than yeasts, Paramecium is a very attractive model. The diversity of the *Paramecium* ciliates is well studied. We focused on four species of *Paramecium* : *P. biaurelia*, *P. sexaurelia*, *P. tetraurelia* and *P. caudatum* as an outgroup (see phylogenetic tree). The three *aurelia* species underwent two rounds of WGDs, WGDX (... years ago) and WGDX (... years ago).

We have shown previous the tight link between gene expression and duplicate retention among *P. caudatum* and *P. tetraurelia*. Highly expressed genes are more retained than low expressed genes ; the regulation of gene expression thus impacts gene retention. *Cis*-regulatory sequences are upstream sequences influencing downstream gene expression.

Transcription Factor Binding Sites, regulatory sequences, regulatory network ?, conserved sequences, studied for a long time. Regulatory sequences ¡-¿ WGD ?

Objective : detect conserved TFBS among various our species -¿ to identify candidates for further studies.

Biological questions : Do gene expression is linked, in *P.*, with specific motifs ? How are TFBS affected by WGD ? Are they conserved among species, is this linked to expression level ? Conserved among each species ? Is there a bias of TFBS usage in certain species ?

Pipeline

Description of the full pipeline

Challenges

Species tree or gene tree ?

Using a phylogenetic footprinting program means we have to use a phylogenetic tree and depending on the phylogenetic tree we are using, the evolutionary signal used in the footprinting is not identical.

The species tree (see figure.) gives us the relation between all considered, accounting for the various splits between species along with WGDs. The problem is that, not all gene families follow this tree. In particular, gene conversion, is heavily involved in shifting genes tree away from species tree : paralogs gene recombine and ...

Motif detection strategy

BigFoot does not output directly identified motifs, instead it produces two files with an alignment of the sequences and associated phylogenetic and alignment scores. The phylogenetic scores is computed, etc. The alignement scores, etc.

Transcription factor binding sites are known to be generally conserved but degenerate on certain positions. For example, ... showed that this motif was conserved ....NN... with two highly variable positions (denoted byN̈; meaning Ä; T̈; C̈ör G̈üsing IUPAC notation). Thus, to seek biologically relevant motifs, we had to take into account that in the middle of motifs, the phylogenetic score could drop on several positions. (Show a phylogenetic score profile ?)

To answer this problem we use a sliding window method of 8 nucleotides in our analysis : for each family, we looked at the scores of 8 nucletodies at the time and slide along the sequences. If the window contained at least 6 bases with scores above our thresholds, we would retain this motif. Then, from this particular region we would try to extend the sequence by adding adjacent nucleotides with good scores.

Measure motif relevance

MEME was shown to have a very high False Positive Rate of the discovery (Ref. needed). That is why many studies combine multiple motif detection tools (Ref. needed).

In our case, MEME is of particular interest as it obtains motifs using a totally different method from BigFoot.

To assess the relevance of motifs found from BigFoot's outputs, we compared them to a well-known motif finding program : MEME. For each family we identified overlapping motifs between MEME and BigFoot. We computed an overlapping index as follows : $0 \leq \frac{nucleotidesincommon}{sizeofthesmallestmotif} \leq 1$, if this index was over 0.9 we would then considered the motif as relevant.

Results

From our orthologs and paralogs gene familes we had 5751 families, only 5008 families were analyzed according to our conditions. Among these 5008 families BigFoot found 811 motifs in 608 families. From these 811 motifs 117 matche

Perspectives

Conservation among species. Major results is that. Divergent resolution of WGD → divergence in motifs ?

After identifying motifs should relate presence/absence of motifs to duplicated genes fate.

Motifs detection should take phylogeny into account for comparative analysis. Not the same value.

Need to improve pipeline for degenerate motifs, for the moment, extract only exact motif in the genome. Need to measure diversity among detected motifs -¿ clustering tools, and suppress rendundancy Implement other motif finding tools to validate results.

Materials and Methods

Acknowledgments

Figure Legends

Tables