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

Since Ohno first hypothesized the influence of Whole-Genome Duplications (WGD), scientists kept showing that a broad number of organisms experienced several rounds of WGD : yeast, insects, Angiosperms, 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 called *paralogs*. WGD can also occur when two closely related species hybridize and form a fertile descendant.

WGDs may be involved in many evolutive radiations as they provide the raw material to explore new evolutionary landscapes. A recent study on the horshoe crab genome underlined that WGDs may be a more common phenomenon than stated. They showed that at least a WGD occurred in this conserved lineage, thus WGDs may not be evolutionary drivers.

Still, it seems that WGDs are widespread along the tree of life and since Ohno numerous models have been developed to explain the retention rate we observe between duplicate genes (reviewed in). Gene Balance Hypothesis is for example a well explored hypothesis in the literature, according to this hypothesis dosage-sensitive gene are more retained because of stoichiometry issues, it would thus explain the over retention of transcription factors and multi-complexes proteins.

To understand the consequences of WGDs we have been studying various *Paramecium* species. *Paramecium* are a Ciliates group (See position on phylogenetic tree). Aury *et al.* showed that at least three round of WGDs occurred in the *Paramecium* genus, two of which occurred in the *aurelia* complex. It is a cryptic species complex of 13 species who are reproductively isolated. Studying the evolutionary fate of duplicate genes among this complex is of great interest.

Recently, several studies unraveled the link between gene expression in gene retention after WGD in *Paramecium*. Genes that are highly expressed are more likely to be retained than genes with a low expression. The COSTEX model proposed by the authors states that gene evolution is more constrained as gene expression is high. It explains well why highly expressed genes are more retained and why they evolve more slowly thant other genes.

Having established this link between gene retention and expression level of the gene, it is normal to try to get better understanding of *Paramecium* regulatory sequences. The compactness of *Paramecium* genomes makes the study of regulatory elements easier than in other eukaryotic species. Indeed, with an average of 300nt long inter-genic regions in *P. tetraurelia* the genome is similar to the one of yeast.

Various strategies to study regulatory elements evolution have been developed ; the idea being that regulatory elements can be conserved along evolution, some tools can detect these more conserved regions and thus output putative motifs. Phylogenetic footprinting is one of the approaches developed when trying to detect motifs among several species, the motif finding process is weighted by the phylogenetical relationships of the given sequences, *id est*, if a motif is found in two closely related species it will have less weight than if it is detected between two distant species.

In this paper we develop a phylogenetic footprinting workflow to study the *cis*-regulatory elements among three *aurelia* species : *P. biaurelia*, *P. sexaurelia*, *P. tetraurelia* and a more basal species *P.*

caudatum. The three *aurelia* experienced two specific rounds of WGD compared to *P. caudatum*, thus when studying orthology and paralogy groups between genes, they can contain up to 13 different genes : 1 from *P. caudatum* and 4 from each *aurelia* species. Using those groups we use our workflow to identify conserved motifs in the upstream regions in each group.

Materials and Methods

Used TranslatorX, PhyML, BigFoot, MEME.
We set up a pipeline to make our analyse (Fig.), the code is available at.

Genomes and Annotation

We used annotation and sequence from our previous analyses for P. and P. etc. (see .) and the additionnal sequence of.

Gene families

Looking at the phylogenetic tree of the *aurelia* species, two WGDs occured at the root and affected three of our species. We have established some gene families from WGD2 using comparison, each family contains a set of orthologous genes between the four *aurelia* species studied, and eventually, the paralogous gene found in each species ; at maximum the families contain 13 different genes. Those families were established previously in our team. For details in the method see

Upstream sequences extraction

We considered only families with at least 4 genes. Then, using our genome assembly and annotation of each species, we extracted upstream regions, upstream of the start codon, from 15nt with a cut-off at 250nt of all genes of the family. If the upstream region of a gene overlapped with another gene we discarded the gene, if the upstream region was less than 15nt long we also discarded the gene. Considering discarded genes, we kept only families with at least 4 members in our datasets.

Coding Sequences extraction and alignment

Phylogenetic footprinting requires a phylogenetic tree to weigh the evolutionary signal of given motifs. A motif conserved between two close species will have less importance than a motif conserved in two distant species. Because we are focusing on the conservation of upstream sequences we chose not to use them to avoid circularity. Instead,corresponding coding sequences (CDS) were extracted and used to model phylogenetic trees for each family. We preferred to have a gene tree over species tree, to avoid eventual inconsistencies because of gene conversion (ref. needed). See Challenges section for explanations on the use of gene tree over species tree.

CDSs in each family were aligned using TranslatorX (ref. needed) a protein-guided alignment software. The Maximum Likelihood (ML) tree was then computed using PhyML (ref. needed) default parameters.

Phylogenetic footprinting

We used a phylogenetic footprinting software BigFoot (ref. needed) to detect highly conserved motifs in upstream sequences. We used 10000 burn-in cycles and 20000 cycles with a sampling rate of 1000 for the Hidden Monte-Carlo Markov Chain (HMCMC) process. BigFoot aligns the given sequences with gaps and tries to identify conserved and non-conserved regions ; it models the evolution of those regions along the phylogenetic tree assuming conserved regions evolve more slowly than non-conserved ones. At the

end of the analysis BigFoot outputs an alignment of sequences used to identify slow and fast evolving regions as well as, for each nucleotide in the alignment, the posterior probability of the alignment, higher values show higher confidence in the alignment, and the phylogenetic footprinting result, higher values indicating higher posterior probability of purifying selection.

Using these scores we detected motifs of at least 6 nucleotides long, alignment score over 0.8 and phylogenetic score over 0.9. Because of the known biological nature of Transcription Factor Binding Sites, we allowed for a "gap" in motifs of 2 nucleotide, so that the scores could drop under the thresholds in those gaps. For detail method of the motif detection using BigFoot results, see Challenges section.

Comparison with MEME

To check our predictions and assess the conservation of found motifs, we compared motifs prediction with those of MEME, a widely used *ab initio* motif finding tools. It searches for statistically significant motifs, with a gapless, local multialignment system.

Motif classification and data analysis

All the analyses were produced using R, scatterplots and graphs were produced using the R package ggplot2.

Challenges

During the pipeline development several methods questions were raised.

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. Because of the successive round of WGDs there are several fates possible for pair of duplicate after the first WGD. Some of these genes may cluster together in the same leaves, if the pairs diverge between species ; another possible outcome is the subfunctionnalization of each gene before the second WGD, meaning before the speciation of the *aurelia* complex, thus genes from different species would cluster together in a phylogenetic tree ; or even a combination of the previous outcome and gene conversion, leading paralogs to recombine in a copy-paste way, changing dramatically the gene tree.

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 alignment 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 by$\ddot{\text{N}}$; meaning $\ddot{\text{A}}$; $\ddot{\text{T}}$; $\ddot{\text{C}}$ör $\ddot{\text{G}}$üsing 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, before rising again.

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 nucleotides 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 gene families had at least 4 genes. Among these 5008 families BigFoot found 811 motifs in 608 families. From these 811 motifs 117 matches with motifs found by MEME.

Perspectives

Conservation among species. Major results is that. Divergent resolution of WGD $\rightarrow$ 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.

Acknowledgments

I would to thank my advisor Jean-François Goût for his patient and constant support, Michael Lynch for having me in his lab. More broadly, I would like to thank the whole Lynch lab team, for great scientific and non-scientific discussions.

References

Figure Legends

Tables

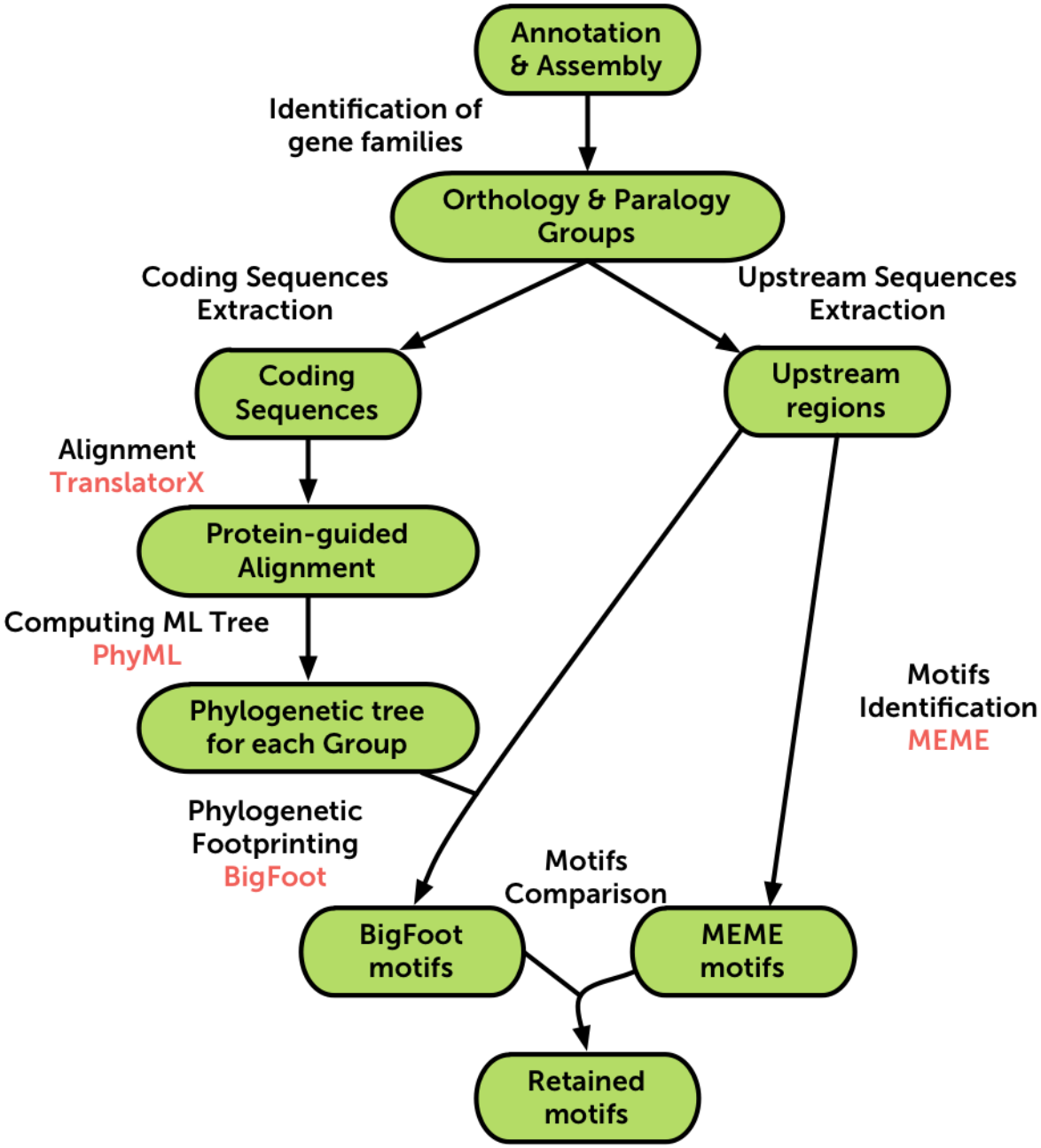

Figure 1. Flow chart of the whole pipeline. From the genome assembly and annotation of the four species, we used orthology and paralogy groups from (Ref. needed). For each of these families we extracted the coding sequences (CDS) as well as upstream regions from the start codon. We built a maximum likelihood phylogenetic tree using PhyML on pre-aligned CDS with TranslatorX. On the one hand we computed the first fifth motifs of size of at least 4 nt using MEME on upstream regions, while on the other hand using both the tree and the upstream sequences were used to detect motifs with a phylogenetic footprinting approach using BigFoot. We then retained only conserved motifs between MEME and BigFoot.