Transcription factor binding sites detection in *Paramecium*

Matthias Grenié [*], **Jean-François Goût** [†] **and Michael Lynch** [†]

[*]ENS de Lyon, Département de Biologie, and [†]Indiana University, Biology Department, Lynch Lab

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed vehicula metus sapien. Suspendisse pulvinar, felis ut hendrerit aliquet, dui nisi bibendum erat, fermentum mattis enim nibh id arcu. Vestibulum ultrices eros sed odio tincidunt bibendum. Pellentesque fermentum ante vel nulla commodo fermentum. Vestibulum in augue sit amet libero viverra accumsan eu at magna. Sed at ligula quis nibh pharetra facilisis non eu libero. Suspendisse non quam sit amet massa luctus interdum sit amet in purus. Integer id orci elit, vitae sollicitudin lectus.

Keyword1 | Keyword2 | Keyword3

Abbreviations: SAM, self-assembled monolayer ; OTS, octadecyltrichlorosilane

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*. As cis-regulatory sequences are broadly

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 ?

Materials and Methods

developped a whole pipeline (show simple pipeline graph)

— Families upstream sequences extraction
— CDSs extraction and alignment
— CDSs phylogenetic tree
— BigFoot identification (explanation of phylogenetic score and alignment score)
— MEME research
— Comparison MEME and BigFoot
— Identification of given motif in species genome
— Correlation between motifs and expression levels
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 analysis 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. We identified 5781 families.

Upstream sequences extraction.

We considered only families with at least 4 genes. Then, using our assembly and annotation of each species genome, we extracted upstream regions 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. Under these conditions, we extracted genes from 5008 families.

Coding Sequences extraction and alignment.

Phylogenetic footprinting requires a phylogenetic tree when detecting motifs, to weigh the phylogenetical 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 prefered to have a gene tree over species tree, to avoid eventual inconsistencies because of gene conversion (ref. needed).

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.

Reserved for Publication Footnotes

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, the stochastic process of the HMCMC let BigFoot refines its alignement. BigFoot assigns to each nucleotide an alignment score, which represents the confidence in the alignment, and a prediction score, measuring the p̈hylogenetic signalöf the given nucleotide, **i.e.**, the more conserved the nucleotide is, according to the phylogenetic tree, the better the score.

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 g̈apïn motifs of 2 nucleotide, so that the scores could drop under the thresholds in those gaps.

Among the previous 5008 families, we identified 1060 motifs in 735 families.

Comparison with MEME.

To confirm our phylogenetic footprinting findings, used MEME (ref. needed) to analyze the motifs in each family. MEME use a statistical process to find motifs. Usually, among a set of given

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.

Perspectives

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

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, etc. Implement other motif finding tools to validate results.

ACKNOWLEDGMENTS. Thanks to M. Lynch and J.-F. Goût. Funding by NSF.

1. M. Belkin and P. Niyogi, Using manifold structure for partially labelled classification, Advances in NIPS, 15 (2003).
2. P. Bérard, G. Besson, and S. Gallot, Embedding Riemannian manifolds by their heat kernel, Geom. and Fun. Anal., 4 (1994), pp. 374–398.
3. R.R. Coifman and S. Lafon, Diffusion maps, Appl. Comp. Harm. Anal., 21 (2006), pp. 5–30.
4. R.R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, Geometric diffusions as a tool for harmonic analysis and structure definition of data. Part I : Diffusion maps, Proc. of Nat. Acad. Sci., (2005), pp. 7426–7431.
5. P. Das, M. Moll, H. Stamati, L. Kavraki, and C. Clementi, Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction, P.N.A.S., 103 (2006), pp. 9885–9890.
6. D. Donoho and C. Grimes, Hessian eigenmaps : new locally linear embedding techniques for high-dimensional data, Proceedings of the National Academy of Sciences, 100 (2003), pp. 5591–5596.
7. D. L. Donoho and C. Grimes, When does isomap recover natural parameterization of families of articulated images ?, Tech. Report Tech. Rep. 2002-27, Department of Statistics, Stanford University, August 2002.
8. M. Grüter and K.-O. Widman, The Green function for uniformly elliptic equations, Man. Math., 37 (1982), pp. 303–342.
9. R. Hempel, L. Seco, and B. Simon, The essential spectrum of neumann laplacians on some bounded singular domains, 1991.
10. Kadison, R. V. and Singer, I. M. (1959) Extensions of pure states, Amer. J. Math. 81, 383-400.
11. Anderson, J. (1981) A conjecture concerning the pure states of $B(H)$ and a related theorem. in Topics in Modern Operator Theory, Birkhaüser, pp. 27-43.
12. Anderson, J. (1979) Extreme points in sets of positive linear maps on $B(H)$. J. Funct. Anal. 31, 195-217.
13. Anderson, J. (1979) Pathology in the Calkin algebra. J. Operator Theory 2, 159-167.
14. Johnson, B. E. and Parrott, S. K. (1972) Operators commuting with a von Neumann algebra modulo the set of compact operators. J. Funct. Anal. 11, 39-61.
15. Akemann, C. and Weaver, N. (2004) Consistency of a counterexample to Naimark's problem. Proc. Nat. Acad. Sci. USA 101, 7522-7525.
16. J. Tenenbaum, V. de Silva, and J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science, 290 (2000), pp. 2319–2323.

17. Z. Zhang and H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignement, Tech. Report CSE-02-019, Department of computer science and engineering, Pennsylvania State University, 2002.

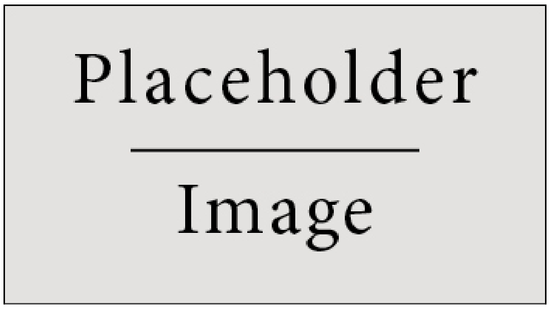

FIGURE **1.** Figure caption

TABLE **1.** **Table caption**

Treatments	**Response 1**	**Response 2**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296