THESE DE DOCTORAT DE

L'UNIVERSITE D'EVRY – VAL D'ESSONNE

2008

David Sergio Armisén Giménez

LES GENES UNIQUES CHEZ LES PLANTES : CARACTERISTIQUES, EVOLUTION ET PROMOTEURS

Composition du Jury :

Dr. Richard COOKE	**Rapporteur**
Dr. Bernard LABEDAN	**Rapporteur**
Dr. Jean-Charles LEPLE	**Examinateur**
Dr. Jeroen RAES	**Examinateur**
Pr. Francis QUETIER	**Examinateur**
Dr. Sébastien AUBOURG	**Directeur de thèse**

Acknowledgements

First of all , I would like to thank Mr. Alain Lecharny who gave me the oportunity to do my thesis in his team and who guided me along these three years. Your words have been always wiseful which made me think it was a quality of Breton people, however I learned a month ago that in fact you was not Breton which means that it should be genuine.

I would also like to thank Mr. Thomas Goujon for his administative effort, with the comity VERT (Versailles-Evry Research Training site), that allowed the INRA center to obtain the Marie Curie fellowships, as well as to Mrs. Hoai-Nam Truong for his work to coordinate the formation of VERT students. I would also like to thank Mr. Michel Caboche who assigned a full-length fellowship to the URGV bioinformatic team which lately allowed me to have my thesis funded.

I acknowledge the members of the jury who have accepted to review my thesis: Dr. Richard Cooke, Dr. Bernard Labedan, Dr. Jean-Charles Leplé, Dr. Jeroen Raes and Pr. Francis Quétier, as well as the members of my thesis committee: Sophie Schbath and Claude Thermes.

Despite it might have been included in the previous paragraph, I would like to thank Mr. Sébastien Aubourg in his own paragraph. You helped and guided me during my thesis but also, once I left my initial shyness, you showed me a completely new world mostly undiscovered for me as the board games. I have really enjoyed your explanations about the different game mechanics and the market news. I only regret don't have started these talks earlier.

I would also like to thank all the members of the team, specially to Véronique Brunaud who's always looking at the bright side of the life, to Jean-Philippe Tamby who gives me his helpful advices to correct this manuscript, to Marie-Laure Martin-Magniette for her inestimable help in statistics and for remain me that I had to think on my conclusions and perspective during ECCB at Cagliari ;-) , and, lastly, to Philippe Grevet who has probed to have a holy patience with me each time that I had a problem.

I want to thank my former colleagues for their support, aid and free french lessons. To Magalie Leveugle ('Mygalie') for all the jokes and funny moments that we spent before she leaves, those were really 'the good days', no pressure and good mood. To Yannick de Oliveira who showed me that despite don´t talking a word in portuguese still have their football skills, I am sorry you couldn't stay longer but I would never forget you, Nicolas, Redouane and Maher, if I learned french

in a first instance was thanks to you. To Clémence Bruyère with whom I shared many lunches and jokes and who was the last one of the three to depart, giving me the last boost.

I would also like to thank Virgnie Bernard with whom I have had many chats, in this case is me who is leaving first, so I am the one encouraging you: keep going, you'll succed and it'll worth it.

I have also a special thought to Jeremy, Mathieu, Véronique, Gwenaëlle, Cécile, Julie and Imen, and all the people from G1 and G2 that I have met, all in their way encouraged me along these three years and made my stay here more comfortable.

Marco y Silvia, los 'otros españoles' de la unidad, gracias por ser como sois, de verdad. Espero que podais ver cumplidos vuestros sueños. Forza azzurri!

Itziar, te conocí el primer día que llegué y te fuiste solo un mes más tarde. Demasiado poco tiempo como para entablar una amistad duradera. Por suerte tú te encargaste de arreglar eso volviendo en varias ocasiones. Cada vez que tú te lamentabas de dejar Málaga yo me alegraba por que volvias. Siento ser tan egoísta pero tu compañía durante la ¿mitad? de mi tésis me ha sido de gran ayuda. Aún nos quedan pendientes unos 'Carcassonnes' pero espero que tengamos tiempo para ellos en un futuro, sea donde sea que el destino nos lleve, tú siempre serás bienvenida.

Ana, mi mayor apoyo y alegria durante estos tres años de tesis. No se donde nos llevara la vida pero se que a tu lado valdra la pena.

A mi hermano Raúl, que ha sido el que peor ha llevado la distancia aunque también el que más ha hecho por reducirla con sus 'interminables' llamadas ;-) Ahora eres tú el que debe emprender su camino y espero que lo hagas como y donde tu quieras. Siempre tendras mi apoyo y estare orgulloso de tí.

Por último quiero acabar agradeciendo a mis padres por haberme dado todo su esfuerzo y dedicación durante tantos años. Aunque estos tres años hayan estado marcados por más de mil kilometros de distancia si hoy puedo escribir estas frases es grácias a vosotros y a lo mejor que me podíais haber dado, la educación. Nunca podre agradeceroslo lo suficente.

We shall go on to the end
whatever the cost may be.
We shall never surrender.
-Sir Winston Churchill-

"Forty-two! Yelled Loonquawl. "Is that all you've got to show for seven and half million year's work?"
"I checked it very thoroughly," said the computer," and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."
"But it was the Great Question! The Ultimate Question of Life, the Universe and Everything," howled Loonquawl.
"Yes,' said Deep Thought with the air of one who suffer fools gladly, "but what actually *is* it?"
-The Hitchhiker's guide to the galaxy-
-Douglas Adams, 1980-

Page

CHAPTER I

INTRODUCTION

I. Introduction

I.A. These basis

Nowadays a number of eukaryotic genomes have been sequenced (23 complete genomes including 3 land plants) or are being sequenced (239 genomes at the assembly step including 11 plants and 223 in progress including 37 plants) (http://www.ncbi.nlm.nih.gov/Genomes). However, it stands to reason that the number of sequenced species will remain, for a long time, extremely low compared with the huge number of the different species that inhabit our planet (about 400,000 plant species have been described and hundreds of new ones are found each year; http://www.bgci.org). In fact, due to the costs and amounts of work necessary to sequence a single eukaryotic genome, up to now, the sequenced plant genomes have been carefully selected on the basis of a number of characteristics including their genome size, position of the species in the evolution tree of the plant kingdom, molecular tools availability for that species and evolutionary proximity with plants of present interest. This model system approach is the key for an efficient acquisition of knowledge on plants of economical interest in a second step. The efficiency of concentrating the international effort on genome models is astonishingly highlighted by the fact that *Arabidopsis thaliana*, the most developed reference species for plant biology has had a significant impact even on human health research (Jones *et al.*, 2008).

Acquisition of a complete genome sequence opens the way to the structural and functional annotations *i.e.* finding information in the sequence. As well as for sequencing, many automatic tools, necessary to cope with an increasing number of sequenced genomes, have been developed to help on sequence annotation. One of the most powerful approaches to help on annotation is knowledge transfer from model species to species of interest. This technique consists in establishing relationships between genes in both species by sequence comparisons in such a way that a biochemical function of a gene characterized in a model plant can be inferred to a gene in the species of interest. Central to the approach is thus the possibility to define reliable evolutionary relationships between genes in two different species. Despite a large effort to develop methods, the search for evolutionary relationships between genes from different species is not trivial and studies at the genome scale always provide us with genes with nearly identical levels of sequence similarities with several genes and for which it is difficult to assign only one unambiguous evolutionary link. It is now generally accepted that all the sequenced species present a variable number of species specific genes sharing no sequence similarity with genes in other species. Nevertheless, it is never possible to rule out the possibility that these *'*species specific*'* genes have

homology links hidden by a too high divergence. We have to keep in mind that (i) gene function is performed by a 3D protein and such structure would not need a sequence conservation farther than in a restricted number of key domains and (ii) in the case of genes with several homologous sequences, *i.e.* belonging to large gene families, knowledge transfer beyond the biochemical level often needs more information than sequences.

I have realized my thesis in the bioinformatics team at URGV (Unité de Recherche en Génomique Végétale, http://www.versailles.inra.fr/urgv/). The bioinformatics team at URGV was involved in the annotation of the *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative, 2000) and of *Vitis vinifera* genomes (Jaillon *et al.*, 2007). Central to the different approaches used is the search for paralogous and orthologous links. The laboratory has coordinated a re-annotation project, named GeneFarm, involving 20 laboratories and using consequences of paralogy to speed up annotations (Aubourg *et al.*, 2005). More recently, the URGV-CATMA transcriptomic resource (Gagnot *et al.*, 2008) has been used to identify novel functional genes in *A. thaliana* that were not found by the AGI (Aubourg *et al.*, 2007). In this context, I asked if comparisons of promoters could help to decide on the orthologous link between two genes from *A. thaliana* and *O. sativa*. In other words, I wanted to show if a foot printing approach might be applied to decide the orthologous relationships among a set of paralogous genes. Due to the small number of sequenced plant genomes, the phylogenetic approach is often difficult due to the numerous paralogous for each pair of orthologous genes. To avoid this problem in a first term, I started from the hypothesis that genes without paralogs but with only one homolog in both the *A. thaliana* and *O. sativa* genomes could be orthologs with closer functional conservation than orthologs belonging to families of paralogs in one or the two genomes. To test this hypothesis, I first clustered and characterized these genes to see if:

(1) Genes without paralogs in a plant genome share specific characteristics that could be due to the selective pressure responsible of their uniqueness?

(2) Unique and species specific genes, *i.e.* genes without homologs in another species and without paralogs, share specific characteristics different from unique genes with homologs?

Second, through an exhaustive study of the promoters, I asked if pairs of promoters of *'*orthologous*'* genes unique in both species share more putative transcription factor binding sites than other pairs of homologs which can be used to define true orthologs among diverse paralogs. This manuscript presents the results of these two steps in two different chapters.

I.B. What are unique genes?

The definition of unique genes is variable depending of the point of view employed. There are two definitions for *'*unique genes*'*:

(1) A sequence-based definition: 'Unique gene' has been employed to define genes with no sequence similarities with other genes. In the frame of this definition, a 'unique gene*'* has no similarity with any other gene in the same genome even if it may have the same function as another gene. Note that sequence similarity is dependent on the measurement of distance between two sequences but not directly related to evolution as it is the case for orthology.

(2) A function-based definition: Based on experimental and functional information, the definition of 'unique gene' has been linked not to the sequence but to the function. Thus, in a genome, a unique gene may have some sequence similarity with several other genes but it may be the only one playing a given function. This definition of 'unique genes' is ambiguous as it could be referred to both biological or biochemical functions. As for the first definition it is independent of any evolutionary information.

In this work we use the sequence-based definition of 'unique genes' and, operationally, we used sequence comparisons to find them into genomes.

Therefore, unique genes are defined inside a genome but two different types of unique genes exist in regards of other genomes. First, there are genes unique in a given genome but that have homologs in genomes from different species. Second, there are unique genes that are specific of a genome, *i.e.* do not exhibit similarities with genes in other genomes. These genes are generally called *'*orphan*'* genes (Fukuchi and Nishikawa, 2003) despite in some cases *'*orphan genes*'* refer to genes with no function that could be inferred from present knowledge (orphan of function) (Domazet-Loso and Tautz, 2003). In this thesis we have considered both types of unique genes, the genes unique in a species and the orphan genes, because we assumed that their comparison could shed light on the evolution of these two types of genes.

But why are these genes so interesting? From an evolutionary point of view, the existence of unique genes should be discussed in the framework of duplication events that are central for speciation. Many studies and data from sequenced species provided evidence of whole genome duplications (WGD) in many species. At first sight, WGD would imply that presence of unique genes should be rare or absent. However, this is not the case and all the species sequenced so far present a variable number of genes with no sequence similarities with other genes (Labedan and Riley, 1995). Indeed, loss of genes after WGD is a massive process that leads to the presence of unique genes in extant genomes as seen in yeast (Dietrich *et al.*, 2004; Dujon *et al.*, 2004; Kellis *et*

al., 2004). Therefore, the analyses of the characteristics and the conservation of unique genes in plant are interesting from a functional point of view as they can provide an insight into unique gene functions that, in some cases, could be important for speciation. Consequently, unique genes are a particular group of (ancient duplicated) genes that could provide some interesting insights on species radiation and evolution.

I.C. Origin of unique genes

I.C.1. Unique genes and whole genome duplication

The existence of unique genes is highly affected by regional and genome duplications since the first consequence of duplication is to eliminate sequence singularity. Question arisen then about the kind and the number of duplication events that unique genes have undergone. First duplication studies considered duplications as local and segmental events involving a variable number of genes. Such duplication mechanisms would imply that the origin of unique genes could be simply due to the fact that, during evolution, some genes were not involved in those duplication events. Thus, the singularity of some genes in extant genomes could be due to randomness. However, further studies in diverse species showed an accumulation of duplications leading to a situation where the most parsimonious model to explain such duplication accumulation was whole genome duplication (WGD). In such a way, in 1970, Ohno proposed that gnathostome vertebrates have underwent two rounds (2R) of tetraploidy events close to their radiation and settled the bases that most of current genes came from ancient duplications (Ohno, 1970). His suggestion that whole genome duplications were an important step in vertebrate evolution was lately supported by the fact that often for a single copy gene in *Drosophila* it is likely to find four paralogs in human (Spring, 1997) and by the number of genes in the genomes of three invertebrate species belonging to divergent phyla (Dehal and Boore, 2005). Simmen *et al.* (1998) estimated the number of genes in invertebrates to range between 11,000 and 19,000, *i.e.* about one quarter of the 50,000 human genes estimated by Ohno (1970). All these data supported the idea of a one-to-four ratio between the numbers of genes encoding protein of invertebrates versus vertebrates which would imply that no ancient vertebrate genes present before the radiation of vertebrates from invertebrates should be found in a single copy nowadays. However, more recent estimations of the human genome reduced the expected number of genes encoding proteins to about 35,000 (Roest Crollius *et al.*, 2000) and this number has been then confirmed by the annotation of the genome sequence (Lander *et al.*, 2001). This reduction in the human gene number does not refute the 2R hypothesis of two rounds of tetraploidy in

vertebrates after radiation from invertebrates. It rather suggests that not all duplicated genes could have survived to evolution. It also put forward as a possibility that the origin of unique genes could be more ancient than previously expected.

Ohno's hypothesis of whole genome duplications along evolution has been validated not only in gnathostome but also in other phyla when more data has been available. For instance, evidences for an ancient whole genome duplication have also been found in the genome sequence of *Saccharomyces cerevisiae* (Wolfe and Shields, 1997), *Arabidopsis thaliana* (The *Arabidopsis* Genome Initiative, 2000), *Oryza sativa* (Yu *et al.*, 2002), *Vitis vinifera* (Jaillon *et al.*, 2007), *Populus trichocarpa* (Tuskan *et al.*, 2006) and *Paramecium tetraurelia* (Aury *et al.*, 2006). Data availability has arisen also another interesting question about the number of WGD events along evolution of diverse species. The origin of this question is that despite firstly postulated 2R hypothesis has been confirmed in various species (Dehal and Boore, 2005), other species as *Paramecium tetraurelia* present evidences to have undergone at least three WGD events during its evolution, and even a fourth one was also probable despite not clear (Aury *et al.*, 2006). In plants, similar duplication evidences from the sequence of the *Arabidopsis thaliana* genome also suggest that it has been duplicated three times during the past 250 million years (Simillion *et al.*, 2002; Bowers *et al.*, 2003; Tuskan *et al.*, 2006; Jaillon *et al.*, 2007).

Therefore, there are evidences that whole genome duplications have extensively occurred in a wide range of plant species at least two times along evolution. How and why have unique genes reverted to single status after these repeated duplications is a question of particular interest which can be helpful to potentially identify genes of functional interest. Additionally, we should not forget that not all unique genes may have an ancient origin but that many of them may have been formed in a given genome after the last whole genome duplication. From an evolutionary point of view, these new genes could provide precious information about species evolution since it is possible that they played an important role in the speciation event.

I.C.2. When the last duplication of unique genes occurred?

Although most unique genes have been duplicated during a defined period of time because of whole duplication events, it is also possible that some of them have arisen after the last WGD undergone by this species. In both cases calculation of the last whole genome duplication age is crucial because elapsed time would show the time available to erase one duplicate as well as the time available for the formation of new genes. The age of duplication in one species may be directly calculated from the study of duplicated gene sequences or could be indirectly calculated from the

divergence time between closely related species already presenting such duplication. In both cases, the age of the duplication event can highly vary depending on the calculation method used: molecular clock and molecular phylogeny.

I.C.2.a. Molecular clock

Methods based on the molecular clock are the most widely used. They allow a direct estimation of the WGD age. The method relies on counting the number of non-synonymous substitutions between proteins coded by duplicated genes to calculate a substitution rate. The time elapsed since gene duplication is then calculated based on previous estimations of substitution rates per million of year on different species from fossil records. Using this approach, the *Arabidopsis thaliana* non-synonymous substitution rate between duplicated gene pairs provided an estimation date for a whole genome duplication of 65 Myr from today (Lynch and Conery, 2000). Such estimation contrasts with other results which placed ancestral tetraploid about 112 MYA using non-synonymous substitution rate analysis between *Arabidopsis thaliana* and tomato (Ku *et al.*, 2000). Why are these estimations so different? In fact, despite being a widely accepted method, molecular clock methods are based on many assumptions that can directly affect the results.

First, it is necessary to have a correct phylogeny to estimate the divergence dates and it is not always an easy task. Second, it is necessary to have a good model of how substitution rates have evolved along evolution. As suggested by Ohno, right after duplication one of the copies may get free from selection pressure which could accelerate the evolution rate and, therefore, increase the estimated divergence rate. Similarly, natural selection could have acted on different moments of gene evolution since duplication modifying the evolution rate by increasing it, as shown for genes related to the immune system (Liu *et al.*, 2008), or decreasing it, as shown for genes coding for proteins with critical functions (Krasowski *et al.*, 2005). Lastly, substitution rates per million of years are estimated from fossil records used to date divergence between different species but which have their own age estimation biases.

Other important limitations common to all molecular clock methods are population size and generation time (Ayala, 1999). On small populations, molecular clock will accelerate as the time needed to fix a mutation along population individuals would be reduced in contrast with large populations where individuals would dilute new mutations. Similarly, short generations will shorten the time needed to fix a mutation by accelerating the number of new offspring in a time given, while long generations will slow down the clock accordingly. Other assumptions are specific to other molecular clock models. In a paper reviewing all the available methods (Welch *et al.*, 2005) it

is shown that they mainly vary in the size of the confidence intervals. Additionally, analyses should also consider that a genome like the *A. thaliana* genome has been duplicated three times during the past 250 million years (Simillion *et al.*, 2002; Bowers *et al.*, 2003). Thus, different results are expected if sets of genes with different divergent times are used.

I.C.2.b. Molecular phylogeny

Phylogenetic analyses are based on sequence comparison between different organisms to estimate their evolutionary relationship from the degree of similarity/distance between the sequences. The differences within the sequences are considered to reflect the evolutionary distance between two organisms and a phylogenetic tree is constructed such a way that sequence changes between branches are minimized.

One of the first phylogenetic marker used for sequence comparison was ribosomal RNA (rRNA) which was useful to define for instance the 'crown' radiation of protists (Sogin, 1991), metazoa and fungi sisterhood (Wainright *et al.*, 1993) and the discovery of Cercozoa (Cavalier-Smith and Chao, 2003). After rRNA, other molecules began to be used as phylogenetic markers thanks to the availability of large amounts of molecular data available in many species. Such alternative phylogenetic markers as nuclear and mitochondrial proteins helped to show that eukaryote lineages described by rRNA were in fact the fruit of technique artefacts and that most known eukaryotes could be organized in six major clades (Arisue *et al.*, 2002; Rodriguez-Ezpeleta *et al.*, 2005), an evidence also found by improved rRNA analyses (Nikolaev *et al.*, 2004).

Despite molecular phylogenetic analyses can theoretically overcome several limitations of molecular clock methods, different methods delivered quite different values for the WGD dating. For instance, the last whole genome duplication in *Arabidopsis thaliana* was dated around 14–24 MYA (Yang *et al.*, 1999; Koch *et al.*, 2000; Koch *et al.*, 2001) by some phylogenetic analyses while another one estimated it probably around 30-35 MYA (Ermolaeva *et al.*, 2003).

Molecular phylogenetic methods present some particular problems that could be gathered in three classes: long-branch attraction (Philippe and Laurent, 1998; Susko *et al.*, 2004), saturation of changes in deep phylogenies (Philippe *et al.*, 2000) and horizontal gene transfers (HGT) (Gogarten *et al.*, 2002).

In brief, we can consider that due to their increased accuracy, phylogenetic analyses are the most appropriate methods to date species divergence. However, they require more genomic information about different species than molecular clock. It is precisely the limitation of genomic

data from close species, which prevents phylogenetic analysis to be used more extensively to date whole genome duplications.

As a final point, it has to be considered that regardless of the method used, whole genome duplication events are not the only source of duplication which could introduce biases on duplication age estimation. Indeed, gene duplications have been estimated to occur more frequently than had been generally thought, with a rate sufficient to completely duplicate a typical eukaryote genome (10,000-100,000 genes) every 100 millions of years (Lynch and Conery, 2000; Lynch *et al.*, 2001). This high rate will not only affect the estimation of the age of the whole genome duplications but could also have consequences on duplicated gene survival and evolution that are not the same as consequences of whole genome duplications.

I.C.3. Gene relationships definition

Unique gene categories are transversal to gene classes based on the origin of genes, *i.e.* the different classes of homologs. During the writing of my paper in *BMC Evolutionary Biology* (Chapter II) and of my thesis in general, I have been faced to the necessity to create a short wording representation of the different complex gene relationships observed between unique genes. Precisely this ability to communicate the exact origin of two homologous genes is determinant in order to discuss the evolutionary and functional consequences of the uniqueness.

In the case of unique genes with no homologous counterpart in any other known species, their definition as orphan genes to note this lack of similarity with any known sequences is quite consensual. However, if unique genes have a homologous counterpart somewhere in the living world, definitions become more complex. Orthologs and paralogs were the two firstly terms introduced to talk about origin of homologous sequences regardless of the number of gene copies (Fitch, 1970). Orthologous genes are genes found in two different species and that come from a single common ancestor gene present at the time of speciation event while paralogous genes are originated from duplications within the same species (Table 1). An example of these two situations can be found in Figure 1: Genes 1A and 2A are orthologs as they both come from a common gene, A, in their last common ancestor, 0. At the other hand, 1A and 1A' genes have their common origin in a duplication event and are found in the same species, therefore they are paralogs.

Orthologs and paralogs terms have been used for a long time. The explosion of the availability of sequenced genomes from different species made necessary to sophisticate the terminology in order to take into account more precise evolutionary assumptions. Difficulties arise for example when orthologous genes come from a single common ancestor from older speciation

Orthologs	Homologous genes found in two different species that come from a single common ancestor gene present at the time of speciation.
Paralogs	Homologous genes originated from duplications within the same species.
Co-orthologs	Two or more genes in a species that, due to a duplication event, are all orthologous to one or more genes in another species.
In-paralogs (or sym-paralogs)	Paralogous genes resulting from duplication occurred after speciation event.
Out-paralogs (or allo-paralogs)	Paralogous genes resulting from duplication occurring before speciation event.
Pseudo-orthologs	Out-paralogous genes that due to an asymmetrically lost on each species are identify as orthologous.
Xenologs	Homologous genes that are the consequence of a horizontal gene transfers (HGT).
Syn-orthologs	Co-orthologs containing paralogous relationships.
Pan-orthologs	Co-orthologs containing only orthologous relationships.

Table 1 – Gene relationship definitions

events, or when they have been duplicated before the species radiation. To help with this increase in complexity, new terms for different types of orthologs and paralogs were proposed: co-orthologs, in-paralogs and out-paralogs (Sonnhammer and Koonin, 2002; Koonin, 2005) (Table 1). The co-orthologs term was suggested to define two or more genes in a species that, due to a gene duplication event, are all orthologous to one or more genes in another species. If the gene duplication occurred after a speciation event, the term suggested for defining this species specific gene duplication is in-paralogs (or sym-paralogs). If the gene duplication occurred before a speciation event the duplicated genes are named out-paralogs (or allo-paralogs). Following the previous example from Figure 1, the genes 1A and 1A' are both co-orthologs of gene 2A, as well as in-paralogs between them. Similarly, 5A and 5A' genes are co-orthologs of gene 3A but their relationship in respect of themselves is out-paralogs as they are duplicated genes which have arisen before the speciation event. The problem with these new definitions is the necessity to know the exact phylogeny of the species and the exact moment of the duplication to distinguish between in- and out-paralogs (Table 1). Another problem arises when, after speciation, two ancestral paralogous genes are asymmetrically lost on each species (Figure 1, genes 3A and 4A'). The resulting situation of asymmetric loss of duplicates, named pseudo-orthology, leads to the wrong assignment of an orthology link to paralogous genes when not enough genomes are available to detect the asymmetric loss (Figure 2). Other terms proposed in the literature are even more specific and deal with homologies due to horizontal gene transfers (HGT). Despite their specificity, they can be very helpful to differentiate the origin of the homologous genes and avoid, as in the case of pseudo-

orthologs, wrong evolutionary or functional assumptions. Horizontal gene transfers was firstly predicted by Syvanen in 1985 as a mechanism of gene transmission between different prokaryote species. This horizontal transmission makes possible that one species has one gene inherited from its ancestor, and a paralogous gene acquired from a related species, as genes 6A and 7A in species 6 (Figure 1). These two genes, which would normally be described as paralogs, are in fact pseudo-paralogs discarding possible duplication events after radiation. A more complex situation is possible if the transferred gene does not co-exist with its paralog but completely substitutes it. In this case, or if the receiving species originally has no paralog as genes 7A and 6A' in species 7 (Figure 1), the resulting comparison between both species would result in the detection of two orthologs. Homologous genes obtained in this way are named xenologs and their identification is very important to discard false conserved genes or genes under a high selection pressure (Table 1).

Alternative and update of these terms have been proposed: syn-orthologs instead of co-orthologs to contrast orthology relationships containing paralogs, *e.g.* genes 5A and 5A' to 3A (Figure 1), from those made of strict orthologs named pan-orthologs, *e.g.* genes 1A and 2A (Blair *et al.*, 2005). Despite the introduction of all these ontologies, defining with a single word the exact origin of two homologous genes can be difficult. Therefore, it is still necessary to use complex sentences or at least a composite of words to explain which genes within a group of co-orthologs come from a common ancestor and which had been duplicated lately.

For reading concerns, we have reduced the complexity of such new terms by mainly using the terms 'orthologs', 'paralogs' and 'pan-orthologs', while the rest of terms have been substituted by a complete phrase describing the situation in each case.

I.D. Gain and loss of functions in unique genes and their consequences on evolution

I.D.1. No-, Neo- and Sub-functionalization of unique genes

Natural selection acts on genes through evolution to conserve genes whose functions are beneficial for the fitness of a species. If we hypothetise that unique genes code for unique functions, it is easy to explain why they have been conserved through evolution. However, as explained before, all unique genes have probably been duplicated several times at some point of evolution. What happens to unique genes during duplicated phases? As we will see in this section, duplication

is essential for unique genes not only to acquire novel functions but also to modify or specialize them.

Right after a duplication both copies of a former unique gene would have fully overlapping functions regardless of the duplication origin: WGD, segmental or gene. According to a model firstly proposed by Ohno, this overlap would allow one of the duplicates to retain the ancestral function while the other avoids selection pressure and accumulates mutations (Ohno, 1970). Ohno proposed that these mutations would end up in the acquisition of new functions that would be then conserved by selection pressure. However, Lynch *et al.* analyses confirmed that apparently the most probable situation is an accumulation of deleterious mutations over beneficial mutations (Lynch and Walsh, 1998). Accumulation of deleterious mutations could lead to no-functionalization, *i.e.* the emergence of a non-functional gene, or a pseudogene, that can be completely lost with time. McClintock *et al*. presented evidences of no-functionalization (Figure 3, gene 2A'') in the Zebrafish Hox family (McClintock *et al.*, 2001).

Not all duplicated genes are lost despite the increased probability of deleterious mutations. Some genes may accumulate mutations granting new functions or specificities that would lead to the fixation of both duplicates due to selection pressures. The scenario proposed by Ohno, would be a specific case of this fixation called neo-functionalization. In neo-functionalization, as observed for instance in the formation of the Retinoic Acid Receptors in vertebrates (Escriva *et al.*, 2006), one of the duplicated genes retains the ancestral function while the other one is free to accumulate mutations and evolve (Figure 3, genes 1A and 1A'). However, this situation of perfectly separated roles on the retention of mutations in the two duplicates has been largely argued during the past years. Such discussions lead to a new model where a gene with two distinct functions would perform both functions right after a duplication event but then, during evolution, one of the duplicates would loss or favour one of the functions and get specialized in a single function (Hughes, 1994). As a consequence of this partitioning of functions, both duplicates would be positively selected to maintain them while fully maintaining the original functions. Unlike neo-functionalization, this concept did not involve the gain of new functions for duplicated genes but was rather based onto the assumption of multiple functions performed by the ancestral gene and asymmetrically conserved by duplicated genes. This situation was defined as sub-functionalization or duplication-degeneration-complementation model (DDC) by Force *et al.* (1999). Some of the simplest ways of sub-functionalization were described for duplicated genes with multiple regulatory elements. In these latter cases, one duplicated gene can randomly loss one regulatory element and once this occurred the rest of duplicated genes would be selected to maintain gene dosage by loosing other regulatory elements (Figure 3, genes 2A and 2A'). As a consequence of this

asymmetrical loss, duplicated genes would have different expression patterns in different cell tissues, development stages and/or under different stimuli.

Several examples of sub-functionalization have been described. For instance, Force *et al.* (1999) observed that tetrapods present two engrailed genes: En1 and En2 (Joyner and Martin, 1987); while Zebrafish present four: eng1, eng2, eng3 and eng1b (Ekker *et al.*, 1992; Amores *et al.*, 1998). Phylogenetic analysis showed that eng1/eng1b and eng2/eng3 come from an ancient duplication of en1 and en2 respectively. Despite common origin, expression patterns of eng1/eng1b genes are different. *In situ* hybridization shows eng1 expression in pectoral appendage bud, while eng1b is expressed in a specific set of neurons in the hindbrain/spinal cord. In the opposite, tetrapod en1 gene presents both expression patterns, suggesting that after radiation and a duplication event, sub-functionalization has specialized expression in eng1 and eng1b genes. Other examples of sub-functionalization were described by McClintock *et al*. (2001) in the Zebrafish Hox family and by Li and Noll (1994) in triplicated Drosophila genes: *paired* (prd), *gooseberry* (gsb), and *gooseberry-neuro* (gsbn) which have different developmental roles despite they have conserved the same functions. Therefore these genes should concentrate their differences in cis-regulatory regions to explain their diversified developmental functions. However, the origin of these cis-regulatory differences is not clear as it could be the result of an acquisition of different cis-regulatory regions after duplication (Figure 3, genes 1A and 1A'), as suggested by Li and Noll (1994), or of the loss of different cis-regulatory elements in each duplicates, as suggested by the DDC model (Figure 3, genes 2A and 2A').

Despite the abundant descriptions of neo- and sub-functionalization in the literature, it is recognized that the conservation of duplicated genes along evolution is not frequent as shown by Lynch *et al.* (Lynch and Walsh, 1998). This low frequency in the conservation of duplicates is exemplified by extant genomes. For instance, in the *Arabidopsis thaliana* genome (The *Arabidopsis* Genome Initiative, 2000) only 30% of the duplicated genes have been conserved, which represents 51% of the current genes. In *Saccharomyces cerevisiae*, the percentage of genes that have retained their duplicated copies is even lower, around 13% (Wolfe and Shields, 1997). Nevertheless some extant genomes show a higher conservation of duplicated genes and, in *Paramecium tetraurelia*, only 32% of the proteome correspond to genes that have lost their duplicates that originated from a WGD but have returned to a single state (Aury *et al.*, 2006). The differences may be mainly explained by the date of the last WGD combined with the evolutionary rate in each organism.

Polyploidy is widespread in plants and most of the cultivated plants are polyploids, in particular allopolyploids. Allopolyploidy consists of whole genome duplication through hybridization of two related species and chromosome doubling, which contrast with autopolyploidy

where genome duplication is done between chromosomes of the same species. After duplication, polyploid genomes undergo genomic rearrangements including frequent deletions, as for example the *Ha* locus in wheat (Chantret *et al.*, 2005). Another example of divergence in allopolyploids is the rapid sub-functionalization observed in the analysis of *Gossypium hirsutum* (Adams *et al.*, 2003), an allopolyploid aged around 1-2 MYA (Zhao *et al.*, 1998; Cronn *et al.*, 2002). In the case of *Gossypium hirsutum*, data suggests that some silencing events are epigenetically induced during the allopolyploidization process (Adams *et al.*, 2003). Such epigenetical changes produce a rapid functional divergence between duplicated genes or even the co-suppression of duplicates (Flavell, 1994). All these changes allow the quick adaptation of the plant to WGD reducing the possible deleterious effects of duplication.

I.D.2. Duplication events lead to unique genes loss?

Duplication events are intimately linked to unique gene evolution as they allow the necessary redundancy to obtain novel functions or split the ancestral ones. However, when a unique gene undergoes a duplication event, its 'unique' tag is eliminated as it now present similarity with other gene (remember that our definition of unique genes is based on sequence, not on function). This event would lead to the loss of all unique genes through evolution, but as explained before, duplicated genes do not remain unchanged and evolve in the course of time. We can consider then what are the consequences of this no-, neo- and sub-functionalization on the 'unique' feature.

The easiest case to consider is no-functionalization or pseudogeneization after duplication. The loss of sequence similarity with another gene returns a 'unique' tag to the duplicate gene maintained functional. The cases of neo- and sub-functionalization are more complex and the 'unique' status of the duplicates depends on the type of modifications that have been involved in the differentiation of both duplicates: regulatory modifications or coding sequence changes. In the case of all the examples of neo- and sub-functionalization previously described, all gene modifications were performed at the regulatory level by changing or losing different regulatory motifs in each duplicated gene. Such changes would not affect coding sequences and duplicated genes would lose their 'unique' tag unless some posterior event eliminates or substantially change the coding sequences. However, it is possible that neo- and sub-functionalization are not the consequence of a regulatory modification but a change in coding sequence. This option might eliminate sequence similarity and return each of the genes to a 'unique' status. Nevertheless, frequently, selected sequence changes affect only a minor number of nucleotides that are enough to change the coded amino-acids in a region involved in a molecular interaction or to alter the folding structure. Both

modifications may bring a new function to the protein. In this way, the number of changes necessary to completely remove sequence similarities are high and in most of the cases, genes would form a gene family rather than to become unique genes.

On the contrary, it is possible that one or both of the duplicated genes undergo major changes in their coded proteins that occasionally might reduce sequence similarity to a negligible level. Such changes, mainly due to exon loss, could be found in some cases of sub-functionalization in genes coding for various subunits with different functions. Indeed, the modularity of proteins has been widely used during evolution to generate novel functions (Sonnhammer and Kahn, 1994; Riley and Labedan, 1997; Le Bouder-Langevin *et al.*, 2002; Liang *et al.*, 2002; Bru *et al.*, 2005). As described for the changes in regulatory functions, after the duplication each gene copy could loss one of these subunits specializing their function not at the expression level but at the level of the biochemical function. Other possible major changes in coding sequence such exon shuffling and fusion/splitting process despite altering the coded proteins would probably not vary gene sequence substantially maintaining genes as not 'unique'.

Finally, it should be noted that, as for regulatory modifications, all these sequence changes could promote the acquisition of new functions or the conservation of ancestral ones, but 'unique' tag is independent of this fact. Therefore, genes may be unique even if they have been generated by duplication and have conserved the same function as the ancestor as long as they accumulate enough mutations to be not recognized as similar to another gene at the sequence level.

I.D.3. Duplication fixation or reversion to unique genes?

Unique gene evolution is intimately linked to duplication that allows the transient redundancy necessary to acquire, or loose, novel functions and/or regulatory elements. Nevertheless not all duplicated genes are fixed during evolution, and as shown by Lynch *et al.* (Lynch and Walsh, 1998) generally one duplicate accumulates deleterious mutations and is lost, and only one copy of the original gene is maintained in the genome. Despite such a description can suggest that the fixation of duplicates is a random process, the probability for duplicated genes to be lost or fixed is not the same for every genes and numerous studies have showed that it is highly dependent on the type of duplication and the function of the duplicated gene (Orr, 1996; Zhang and He, 2005; Scannell *et al.*, 2006).

I.D.3.a. Type of duplication and duplicate fixation or loss

The type of event leading to gene duplication is very important for duplicate fixation. On one hand, duplications of chromosome segments or individual genes can destabilize an established gene dosage, particularly when they involve members of a protein complex or of a regulatory network (Wapinski *et al.*, 2007). Such an unbalanced gene dosage can be detrimental for the cell fitness and, thus, will be counter-selected (Birchler *et al.*, 2007). On the other hand, WGD events duplicate not a single but all genes at a time and keep gene dosage unchanged. Consequently, the duplicated genes issued of a WGD event would have a better chance to be conserved. Further mutations affecting the functionality (no-functionalization) of a gene involved in a network or a complex seriously altered by this change will change the pressure direction towards a strong negative selection (Aury *et al.*, 2006). Indeed, it has been observed that the genes involved in complexes have been preferentially retained after genome duplications (Papp *et al.*, 2003; Maere *et al.*, 2005; Blomme *et al.*, 2006) while the genes issued from small-scale duplication events have a low retention rate (Krylov *et al.*, 2003; Papp *et al.*, 2003).

The correlation between Protein-Protein Interactions (PPI) and duplication level has been analysed in *Saccharomyces cerevisiae* (Wagner, 2002; Makino *et al.*, 2006). Results showed that one duplicate tends to have more PPI than the other one and evolves at a slower rate to maintain the original function.

I.D.3.b. Gene function and duplicate fixation or loss

The link between the retention rate of duplicated genes and the mechanism of gene duplication may be considered as an indication of the importance of the gene function onto the retention rate.

Different studies indicated that gene function is an important factor on maintaining duplicated genes. In eukaryotes, genes coding for transcription factors, as well as other genes involved in signal transduction and development have been preferentially retained after genome duplications (Blanc and Wolfe, 2004; Maere *et al.*, 2005). In contrast, these genes have a low retention rate after small-scale duplication events (Krylov *et al.*, 2003; Papp *et al.*, 2003). Wapinski *et al*. analyses in *Ascomycota* phylum (which includes *Saccharomyces*) the genes that encode for peripheral transporters, receptors and cell wall proteins, and genes tend to lose their duplicates more than the genes essential for viability, involved in essential growth processes or residing in the nucleus, nucleolus, mitochondrion, endoplasmic reticulum and Golgi apparatus (Wapinski *et al.*, 2007). In *Arabidopsis thaliana* and *Saccharomyces cerevisiae*, the genes involved in transcription

regulation and signal transduction were preferentially retained in duplicates (Seoighe and Gehring, 2004). However, it has been suggested that such retention of duplicates could be deleterious due to an increase in the metabolic cost necessary to synthesize more proteins (Wagner, 2005).

Thus, both the dosage balance and gene function would explain why after gene duplication the duplicated genes retained after a previous duplication have a higher probability of being maintained duplicated while duplicated singletons are rapidly returned to the single copy status (Seoighe and Gehring, 2004; Chapman *et al.*, 2006). In *Arabidopsis thaliana*, the probabilities for these two events have been estimated to be 0.26 and 0.17 respectively (Seoighe and Gehring, 2004).

I.D.4. Evolutionary consequence of unique genes

Gene duplications can be the base for acquisition of new functions or loss of existing ones, as well as gain or loss of regulatory elements. Gene duplication can also conduct to the split of ancestral functions which would lead to situations of neo- and sub-functionalization. Such situations occur on individual level and it is possible that different populations of the same species accumulated a distinct set of alleles during evolution. While divergent alleles would accurately perform their functions in their respective populations, it is possible that they entail deleterious effects when present in the same genome after crossing. As a consequence of this incompatibility, hybrid offspring of the two populations would have reduced viability or fertility that could even lead to complete reproductive isolation. This situation of divergent evolution and posterior incompatibility is known as the 'Dobzhansky-Muller incompatibilities' or DMIs (Orr, 1996; Coyne and Orr, 2004; Brideau *et al.*, 2006). Thus, it is possible that unique genes have an active role on species isolation due to the accumulation of different mutations either in their coding sequence or cis-regulatory regions. This hypothesis is supported by a comparison of the *Saccharomyces cerevisiae*, *Saccharomyces castellii* and *Candida glabrata* genomes (Scannell *et al.*, 2006). Analysis of the gene conservation after WGD indicates that when one duplicate was lost, often it was the same in the three species. However, a small but significant proportion of duplicated genes indicate a process of reciprocal gene loss. Reciprocal gene loss is a particular form of reciprocal silencing (Werth and Windham, 1991) or divergent resolution (Lynch and Force, 2000; Taylor *et al.*, 2001) where each species, after some time during which duplicates diverged, loose a different duplicate and keep, therefore, different copies (see species 3 and 4 in Figure 1). This reciprocal loss of duplicated genes could create a DMI (Werth and Windham, 1991; Lynch and Force, 2000). This incompatibility hypothesis was reinforced by the observation that in *S. cerevisiae* 40% of duplicates involved in a reciprocal gene loss have an essential function while only 20% of genes

involved in non-reciprocal gene loss are essential (Guldener *et al.*, 2005; Scannell *et al.*, 2006). Similar indications of speciation promoted by reciprocal gene loss after duplications have been reported for *Paramecium tetraurelia* (Aury *et al.*, 2006; Hughes *et al.*, 2007), polyploid plants (Werth and Windham, 1991; Paterson *et al.*, 2004) and fishes (Taylor *et al.*, 2001; Postlethwait *et al.*, 2004).

Different types of hybrid incompatibilities may be produced by inbreeding population with unique genes generated by reciprocal gene loss of ancestral duplicates. Among them are the sterility of hybrids (Schnable and Wise, 1998; Coyne and Orr, 2004; Haerty and Singh, 2006), the hybrid tumour formation (Ahuja, 1965; Joshi, 1972; Phillips and Reid, 1975) and the hybrid necrosis (Orr and Irving, 2001; Yu *et al.*, 2005). It is interesting to note that studies have shown that DMI formation tend to be asymmetrical and incompatibilities caused by gene introgression from species 1 into species 2 are likely to not be observed when gene from species 2 is introgressed in species 1 (Welch, 2004).

In *Saccharomyces*, the relevance of DMIs on speciation is indicated by the high rate of loss of duplicated genes in the time interval between WGD and the first speciation event (Scannell *et al.*, 2006). Other studies placed reciprocal gene loss after speciation as in *Tetraodon nigroviridis* and *Danio rerio* (Semon and Wolfe, 2007).

Despite a single incompatibility could potentially lead to a speciation event, it has been estimated that about 20 to 30 reciprocal gene loss of essential genes are necessary to result in a reproductive isolation by a Dobzhansky–Muller process (Werth and Windham, 1991; Lynch and Force, 2000). Overall, Dobzhansky-Muller incompatibilities are possibly responsible of plant speciation when other classical speciation events, as geographical isolation, are overcome. In this context, unique genes are a particularly interesting source of possible DMIs. Even if they may not be the first reason of speciation, they might maintain or increase the reproductive isolation.

I.E. Promoters of Unique genes

I.E.1. Promoter organization

Before a protein can perform its function, the corresponding gene must be transcribed. The promoter is the DNA sequence located upstream the translation initiation codon (ATG) with its 'core' region around the Transcription Starting Site (TSS). It contains specific motifs recognized by proteins responsible of the transcription, the Transcription Factors (TF). Such motifs, called Transcription Factor Binding Sites (TFBS), bind not only the RNA polymerases (responsible of

DNA transcription) but also a variable number of other transcription regulators. Some conserved elements have been observed in promoter regions of both, prokaryotes and eukaryotes, but most of the motifs and positions were different in each case (Cooper and Hausman, 2007).

I.E.1.a. Prokaryotic promoter

Two conserved features have been defined in the prokaryotic promoters: the presence of a TATA-box, also called Pribnow-box (Pribnow, 1975), at position -10 from the TSS; and a TTGACA sequence at position -35 (Figure 4A). In prokaryotes, it has also been observed that the same promoter may be used to initiate the transcription of adjacent genes thus, producing a single mRNA coding for different proteins. This grouping of genes is called an operon and can be controlled by different transcription factors and even different TSS according to the situation. One of the best-studied operon systems is the *lac* operon from *Escherichia coli* (Jacob *et al.*, 1960; Beckwith, 1967).

I.E.1.b. Eukaryotic promoter

In eukaryotes, the core promoter, *i.e.* the region around TSS, contains a number of conserved motifs higher than in prokaryotic promoters, including (Figure 4B):

- TATA-box: The first core promoter element identified in eukaryotes (Lifton *et al.*, 1978) consists in a TATA-box with the consensus sequence 5‘-TATAWAAG-3’ (Wong and Bateman, 1994; Patikoglou *et al.*, 1999). The TATA-box is located at positions -25 to -30 upstream to the TSS. It is initially recognized by a TATA-binding protein (TBP) in yeast (Buratowski *et al.*, 1988) and subunits of TFIID complex in Drosophila (Hoey *et al.*, 1990) and human (Tanese *et al.*, 1991). The other subunits of TFIID complex are called TBP-associated factors (TAFs) (Dynlacht *et al.*, 1991). Once TBP is bound to a TATA-box, it recruits TFIIA, TFIIB, TFIIF, RNA polymerase II, TFIIE and TFIIH in this order.
- Inr: or Initiation element (Smale and Baltimore, 1989) is a conserved region placed around positions –3 and +5, including the transcription starting site. Unlike TATA-box, its sequence is more degenerated but is also commonly found.
- CAAT-box: A conserved motif close to -80 bases with consensus sequence 5’-GGYCAATCT-3’ (Benoist *et al.*, 1980).

- GC-box: A series of motifs rich in guanine and cytosine nucleotides normally found near the TATA-box.
- Downstream Promoter Element (DPE): A conserved motif necessary for the binding of TFIID that, unlike the other conserved motifs, is located downstream the TSS, around positions +28 to +32 (Kutach and Kadonaga, 2000).

I.E.1.c. Promoter variability

Conserved regions in prokaryotic and eukaryotic promoters should not be considered as compulsory promoter elements. In the case of eukaryotic promoters, for example, the canonical TATA-box (TATAWA) is not essential for transcription and, in fact, is present in only 20% of the yeast genes (Basehoar *et al.*, 2004), in 10% of the human genes (Gershenzon and Ioshikhes, 2005; Kim *et al.*, 2005) and 29% of Arabidopsis genes (Molina and Grotewold, 2005). On the contrary, DPE is generally found in TATA-less promoters but not exclusively (Kutach and Kadonaga, 2000). In other cases, sequences are more degenerated than consensus sequence or the positions do not correspond to those described as usual. It is even possible that both TATA-plus and TATA-less promoter regions are present in the same gene but regulating distinct transcription starting sites (Goodyer *et al.*, 2001). Regardless of the regulation mechanism, gene transcription needs a ~100 bases region for the assembly of RNA polymerase II complex and the positioning of the transcription starting site (Reinberg *et al.*, 1998; Lee and Young, 2000).

The RNA polymerases are responsible for DNA transcription but promoter regions do not only contain binding sites for RNA polymerases. They also contain binding sites for many other Transcription Factors. Indeed, promoter regions may contain many different binding sites for different regulatory elements that interact together to control the transcription rate and time of expression. These regulatory elements may be located far from the TSS (upstream or downstream of it) and can be divided in enhancers, silencers and insulators (Cook, 2003).

Enhancers: Enhancers are regulatory elements that increase the transcription of a gene. Enhancer sequences are more common in eukaryotes than prokaryotes as their transcription control is usually positive and TF activate the transcription (Xu and Hoover, 2001; Arnosti and Kulkarni, 2005).

Silencers: Opposite to enhancers, silencers actually bind TF suppressors that reduce or even prevent gene expression. Expression is therefore interrupted until an inducer binds the TF, changing its shape and breaking the TFBS bind. This mechanism allows rapid

responses to environment changes and is therefore more usual in prokaryotes (Ogbourne and Antalis, 1998).

Insulators: Enhancers and silencers fully functional in the regulation of a gene may be located far from the TSS of this gene. As a consequence, an enhancer or silencer may be located 'close' to another gene and therefore might interact with its promoter. To avoid that an enhancer or silencer can improperly act on other genes in its neighbourhood, other elements called insulators could be found placed between gene promoters to prevent that the specific activation of one gene affects other genes around it (Burgess-Beusse *et al.*, 2002; Brasset and Vaury, 2005).

I.E.2. Phylogenetic footprinting

Phylogenetic footprinting is one of the most powerful approaches of comparative genomics to improve cis-regulatory element detection. Firstly introduced by Tagle *et al.* in 1988, phylogenetic footprinting is based on the higher degree of sequence conservation expected on regulatory motifs across promoters of orthologous genes in different species. In other words, as TFBS are functional units surrounded by non-coding DNA, it is expected that they would be under selective pressure while surrounding DNA would not. Such difference on selective pressure would be therefore detectable upon sequence comparison of various orthologous promoters.

First methods designed to detect motifs were based on multiple alignments of promoter sequences followed by the identification of well-conserved aligned sequences (Duret and Bucher, 1997; Hertz and Stormo, 1999; Hughes *et al.*, 2000). More recent methods rely on detection of best conserved motifs within homologous regions regardless of alignment. They are based on phylogenetic relationship (Blanchette and Tompa, 2003) or clusters of un-gapped conserved sub-sequences (Grad *et al.*, 2004). Other methods include the use of known TFBS to model position-specific weight matrices (PWMs) that assign a score to relative frequencies of each nucleotide at each signal position (Frech *et al.*, 1993). Such PWMs take then advantage of TFBS contained in databases such as TRANSFAC (Matys *et al.*, 2003) or JASPAR (Sandelin *et al.*, 2004) as training sets for posterior motif detection in promoter sequences without the need of alignment or multiple orthologous genes. Regardless of the detection method employed, phylogenetic footprinting has successfully been applied on close species for the discovery of regulatory sequences in animal promoters (Tagle *et al.*, 1988; Leung *et al.*, 2000) as well as in plant promoters (Kaplinsky *et al.*, 2002). The phylogenetic footprinting use is limited by the difficulty to align long non-coding sequences from distantly related species (Clark, 2001). Better results are obtained with closely

related species as shown in a study on the regulatory sequences involved in the regulation of the AGAMOUS gene and conserved in 29 *Brassicaceae* species (Hong *et al.*, 2003). This modification of the phylogenetic footprinting approach has been named *'*phylogenetic shadowing*'* (Boffelli *et al.*, 2003) and has been used to detect weakly conserved ancestral mammalian regulatory sequences by primate sequence comparisons (Wang *et al.*, 2007). With the larger access to plant genome sequences expected in the near future, this approach will be a powerful tool to annotate promoters.

One objective of my thesis was to explore a novel application of the phylogenetic footprinting approach what we called the reversal phylogenetic footprinting. In the reversal phylogenetic footprinting methods we wanted to take advantage of TF motifs conservation to establish or confirm orthology relationships. While sequence modifications on the coding sequence can have different effects on final protein, ranging from no effect to complete disruption of encoded protein, selection pressure would tend to minimize completely deleterious mutations but allow other mutations. At the contrary, gene regulation is a very sensitive mechanism controlled by TFs that bind to small conserved motifs within promoter sequences. Therefore, while a single nucleotide substitution in coding region can slightly change the shape of protein or even have no effect at all, a mutation in a TFBS is more likely to disable TF binding and completely deregulates gene expression. Based on this supposition, with reverse phylogenetic footprinting we aimed to use known conserved motifs of a gene family to search for genes with common origin but with divergent coding sequences.

CHAPTER II

UNIQUE GENES AND ORTHOLOG ANALYSES

This chapter has been published in **BMC Evolutionary Biology** 2008, 8:280 with the title '*Unique genes in plants: specificities and conserved features throughout evolution*' (Armisen D, Lecharny A, Aubourg S). Only small format changes have been performed.

II. Unique genes and ortholog analyses

II.A. Background

The role of gene duplications in evolution was suggested forty years ago (Taylor and Raes, 2004). More recently, complete sequencing of several eukaryotic genomes showed the quantitative importance of duplicated genes (Tekaia and Dujon, 1999; Wapinski *et al.*, 2007). In particular, plant genomes contain a high proportion of duplicated genes and, in several plant gene families, the number of paralogous genes is more than one hundred (The *Arabidopsis* Genome Initiative, 2000; Yu *et al.*, 2002). Frequent gene duplications (Lynch and Conery, 2000), occasional segmental (Koszul *et al.*, 2004), chromosomal and genomic duplications (Holland, 1997; Spring, 1997; Pebusque *et al.*, 1998; Simillion *et al.*, 2002; Blanc and Wolfe, 2004; Paterson *et al.*, 2004) shaped present genomes. The underlying mechanisms indicate that the primary molecular events in gene duplication should affect most of the genes independently of their function. Nevertheless, all characterized genomes include single-copy (unique) genes, *i.e.* genes without apparent homolog in the same genome (Rubin *et al.*, 2000) and, for some of them, without any homolog, even in phylogenetically close relatives (Llorente *et al.*, 2000). Indeed, evolution is not a one-direction process and a high proportion of duplicated genes are rapidly lost (Lynch and Conery, 2000; Lynch and Conery, 2003; Scannell *et al.*, 2007). This definition of unique gene is fully independent of the gene function and is only based on the protein sequence uniqueness in the whole proteome for a considered species. For instance, in the framework of this definition, the bHLH transcription factors, whatever the different functions that might be assigned to each of them, are not considered as unique because they all share sequence similarity and, as such, are thought to have arisen from a common ancestor. In other words, in this paper we define as single-copy or unique gene, a gene coding for a protein without detectable sequence motif or global similiarities with any protein in the same proteome.

Unlike gene duplication, gene loss is not an unspecific mechanism but it is instead influenced by functional selection (Kondrashov *et al.*, 2002; Blanc and Wolfe, 2004). Thus, duplicates that are maintained show a bias toward certain gene functional classes (Maere *et al.*, 2005) or transcriptional level (Seoighe and Wolfe, 1999; Lynch and Conery, 2000; Krylov *et al.*, 2003). Unique genes may also be duplicates that diverged too much to be distinguished now (Gaillardin *et al.*, 2000). With the recent availability of whole plant proteomes it is possible to consider further some questions about the generation and evolution of unique genes in plants. In many evolutionary studies, sound groups of duplicated genes are selected but the genes left apart by

the process are far from being all unique genes. Indeed, the potential adaptive significance of duplicated genes and genomes has received great attention (Lawton-Rauh, 2003; Clauss and Mitchell-Olds, 2004; Gutierrez *et al.*, 2004). It is however more difficult to speculate on the meaning of species- or phylum-specific unique genes mainly because of a critical lack of functional annotation for most of them. Major differences in gene repertoire among species were attributed to proteins with obscure features that lack currently defined motifs or domains (POFs) and are often species- or phylum-specific (Gollery *et al.*, 2006). The definition of POFs (Chothia *et al.*, 2003) relying only on the absence of characterized conserved sequence signatures is thus independent of the existence or absence of paralogs. POFs and unique genes are nevertheless overlapping populations of genes. Hypotheses on the possible origins of POFs include convergent evolution and rapid divergence (Gollery *et al.*, 2006). The question of the origin of unique genes, either purifying selection against duplicates or rapid divergence, remains unsolved. In this study we first established and used stringent criteria in order to identify suitable sets of unique genes present in the extensively known proteomes of *Arabidopsis thaliana* (core eudicotyledons, Brassicaceae) and *Oryza sativa* (Liliopsida, Poaceae), two plants that diverged ~150 million years ago (MYA) (Wolfe *et al.*, 1989; Chaw *et al.*, 2004). Second, we used the intersection between the two sets of unique genes in order to characterize a set of genes conserved as unique in both *A. thaliana* and *O. sativa*, *i.e.* pan-orthologs as defined by Blair *et al.* (Blair *et al.*, 2005). Third, we searched for gene, promoter and protein features shared between all unique genes and/or within pairs of pan-orthologs. Fourth, using the pan-orthologs between *A. thaliana* and *O. sativa*, we searched for their conservation in a green unicellular alga and a moss for which reasonably good proteomes are also available. Within the limits of the proteomes used, we show that several unique genes are species specific but that a significant number are conserved even outside of the green phylum. The clusters of homologous unique genes highly conserved throughout the green phylum globally present specific structural features that indicate a strong purifying selection supporting the orthology links between the conserved unique genes. These conserved unique genes would be important targets for functional studies since it is likely that they perform ancient but not described biological functions.

II.B. Employed methods

II.B.1. Data sources

The complete proteomes were obtained from TAIR (TAIR) for *A. thaliana* (R6), TIGR (TIGR) for *O. sativa* (R3), and JGI (JGI) for *P. patens* and *O. lucimarinus*. For *A. thaliana* and *O.*

sativa, we retrieved data concerning the number of transcripts, the PFAM motifs and the promoter sequences from FLAGdb^{++} (Samson *et al.*, 2004). Expression data were obtained from CATdb (Gagnot *et al.*, 2008) and Genevestigator (Zimmermann *et al.*, 2004).

II.B.2. Unique gene characterization

All the proteins encoded by the nuclear genes of each species were retrieved and those from pseudogenes were removed. To identify genes coding for proteins unique in a genome, three different filters were successively applied to the genes (Figure 5). The first filter used the PFAM resource (Bateman *et al.*, 2004) and was selected based on the fact that proteins with common protein motifs are most often homologs. The detection of PFAM motifs is based on HMM profiles (through the HMMER tool) which are more adapted than simple sequence comparisons for the definition of conserved regions, allowing us to eliminate paralogs. All the proteins without PFAM motifs were saved in a list of candidate unique proteins and those with PFAM motifs were re-filtered to select as candidates only the proteins for which the PFAM is unique in the analysed proteome. Second, the proteins encoded by candidate unique genes were compared against the whole proteome through BLASTp. Indeed, the fact that the PFAM resource does not tag around 30% of *A. thaliana* and *O. sativa* proteins and the risk that the PFAM filter introduces bias in tagging preferentially large proteins is corrected by additional BLAST analyses. Furthermore, we have taken care that our BLASTp parameters allow the detection of similarities between very small proteins: Proteins giving an e-value lower than e-10 with another protein in the same genome were discarded from the unique gene list. Third, the genes giving an e-value between e-5 and e-10 with another sequence were considered as unique genes only if they showed a partial match not larger than 30% of their sequence length (size ratio filter). This cut-off (size ratio filter), based on manual expertise of numerous blast results, permitted us to keep genes with hits too small to be considered as probably good despite the e-value obtained.

II.B.3. Conserved single copy genes

A BLASTp of the unique proteins of each species was launched against a database containing the unique protein sequences from every other species. Pairs of proteins showing an e-value lower than e-10, or up to e-5 but satisfying the condition imposed by to the size ratio filter described above, were classified as conserved between the two species. Conserved proteins were then separated into two groups, the U[1:1] proteins if there was only one positive hit or the U[1:m]

proteins if there were more than one hit. U[1:1] genes characterized in each species were compared to select only reciprocal best hits (RBH) and allowed us to remove some U[1:1] in one species qualified as U[1:m] in other species due to a splitting/fusion process. A second BLASTp was launched with those proteins without any hit against a database containing all the proteins from every other species. Applying again the same e-value and size ratio filter as described above, we clustered them as U[1:m] proteins if they had more than one hit, and as U[1:0] if they had no hit on the other species, *i.e.* the species specific unique proteins.

II.B.4. Genomic organization of unique genes

The limits defining the boundaries of duplicated regions in *A. thaliana* and *O. sativa* genomes were retrieved from TIGR database. The even distribution of each group of unique gene pairs between the chromosomes was tested using a chi-square (χ^2) test with a confidence level of 99.5% (expected value of 14.86 and 26.76 for 4 and 11 degrees of freedom, respectively).

II.B.5. Unique gene and protein features

All the different information about genes and proteins was retrieved from the FLAGdb^{++} database (Samson *et al.*, 2004). Information includes protein lengths, number of exons, intron positions and promoter sequences. Only the genes with CDS fully covered by experimental transcript data were used (17,108 and 15,814 nuclear genes in *A. thaliana* and *O. sativa* respectively). For the analysis of promoter sequences, only genes with at least one cognate transcript covering the regions were studied (14,689 and 17,720 for *A. thaliana* and *O. sativa* respectively). Intron positions were compared after aligning protein sequences with ClustalW (Chenna *et al.*, 2003). Intronic conserved positions included those that diverged by not more than 5 amino acids to take into account minor variability in intron position found in different organisms (Boudet *et al.*, 2001). For promoter analyses, the TSS (Transcription Start Site) was defined as the point where the 5' UTR (minimum size of 50 bases) started and promoter sequences comprised the 1,000 nucleotides upstream from it. Positions of such well-known promoters as the TATA (TATAWA consensus (Lifton *et al.*, 1978)), TELO (AAACCCTAA consensus (Tremousaygue *et al.*, 2003), SORLIP2 (also called motif II: GGCCA consensus (Hudson and Quail, 2003; Tremousaygue *et al.*, 2003)) and CAAT (CCAAT consensus (Bucher and Trifonov, 1988)) boxes in each species were set with a program developed by (Bernard *et al.*, 2006) capable of defining significant TFBS preferential positions in promoter regions avoiding false positives (Yamamoto *et*

al., 2007). If the TSS defines position 1, in *A. thaliana* preferential positions were set at: -40 to -21 for TATA-box; -60 to 140 for TELO-box; -240 to -21 for SORLIP2-box and -160 to -41 for CAAT-box. Similarly, in *O. sativa* TFBS were searched for in the following regions: -40 to -21 for TATA-box; -80 to 180 for TELO-box; -280 to -1 for SORLIP2-box and -200 to -1 for CAAT-box.

II.B.6. At and Os U[1:1] gene expression

We based our estimation of the correlation between U[1:1] gene expression in *A. thaliana* and *O. sativa* on EST/cDNA resources. The numbers of associated transcripts of each gene were normalized and logarithmically transformed for comparisons purposes. Normalization avoided biases caused by both the number of transcripts available and the different number of genes for each species. The normalization established an equivalence of 1.56 transcripts in *O. sativa* for one transcript in *A. thaliana*. Comparisons of observed values were made against values from 100 random samples of 937 nuclear gene pairs. To avoid sampling biases due to genes with none or very few transcripts, we only considered the gene pairs with at least 30 cognate transcripts for each member. Furthermore, the random samples only contained protein pairs with a maximum size difference of 20 amino acids between the two members.

II.B.7. Phylogenetic and functional analyses

The phylogenetic evolution of unique genes was analysed from *Ostreococcus lucimarinus* (Prasinophyceae) to *Arabidopsis thaliana* including *Physcomitrella patens* (Funariaceae) and *Oryza sativa*. With the unique gene characterization method (described above), we systematically searched for unique proteins in the available proteomes of the four species studied. Once obtained, we used them in a BLASTp search to look for *O. lucimarinus* unique proteins with a pan-ortholog on each branch of evolution (Figure 5). By this way, we first constructed U[1:1] protein pairs between *O. lucimarinus* and *P. patens*. After, *O. lucimarinus* U[1:1] proteins were used in a new BLASTp comparison against *O. sativa* unique proteins to found U[1:1:1] proteins, and so on until the characterization of the U[1:1:1:1] proteins. Similar protocol was performed starting from *A. thaliana* unique proteins and looking for their conservation on each node of the tree. Both lists of U[1:1:1:1] genes, one per sense, were crossed to eliminate inconsistencies and obtain a final list of 192 U[1:1:1:1] proteins. We calculated the expected conserved number of U[1:1:1:1] genes from *O. lucimarinus* to *A. thaliana* under the no selection pressure hypothesis. The expected number of U[1:1:1:1]genes assuming random conservation was calculated as the number of *O. lucimarinus*

genes (7,618) multiplied by the combined probability of a gene being conserved as unique in *O. lucimarinus* (35.32%), *P. patens* (23.22%), *O. sativa* (13.88%) and *A. thaliana* (9.58%). The expected number of U[1:1:1:1] genes by random conservation would be 8.38 genes. For each species pair permutation, unique genes were aligned with their corresponding ortholog using ClustalW, and the synonymous and non-synonymous substitution rates (dN and dS) were calculated using the Codeml program of the PAML package (Yang, 2007). The protein pairs considered as too divergent by Codeml were nevertheless taken into account in the median dN/dS calculation. For comparison, dN and dS values were also calculated with the same method from a set of 7,551 orthologous proteins predicted by the RBH method using the 3 proteomes of *A. thaliana*, *O. sativa* and *V. vinifera*. Conservation of U[1:1:1:1] genes in other species and functional information were retrieved from the results of BLASTp against the Uniprot database, with a limit e-value of e-10. Comparisons were done against the results of 100 random samples of 200 nuclear genes. The subcellular localization of *A. thaliana* proteins deduced from predictions of signal sequences (based on PSORT, PREDOTAR and CHLOROP software) were recovered from the FLAGdb^{++} database (Samson *et al.*, 2004). The presence of cis-regulatory motifs within promoters was searched with the protocol previously described (Bernard *et al.*, 2006).

II.C. Unique genes analysis

II.C.1. How many unique genes in *Arabidopsis thaliana* and *Oryza sativa*?

With the scope to search for possible evidence of particular features of the unique proteins, our method should be stringent enough to deliver a minimum level of false positives. To achieve this objective we used a protocol that mixed detection of conserved motifs (through the PFAM library (Bateman *et al.*, 2004)), and local sequence alignments (BLASTp) taking into account the relative length of the conserved regions. *A. thaliana* and *O. sativa* were the first two plants with a whole genome sequenced and annotated (The *Arabidopsis* Genome Initiative, 2000; Yu *et al.*, 2002). The corresponding proteins have been used separately to run our protocol for each species (Figure 5). In a first stringent step, we removed 18,274 *A. thaliana* and 28,482 *O. sativa* proteins tagged with the same PFAM motifs. In a second step, remaining proteins were used as query sequence in a BLASTp search (Altschul *et al.*, 1990) against their corresponding proteome. Proteins that returned a hit with an e-value higher than e-10 were filtered on the basis of size ratio value of the best alignment between both proteins. This third step led us to consider as homologs, and thus

not unique, proteins sharing low sequence similarities that are distributed on more than 30% of the full-length protein. Following this pipeline, we found 2,570 unique proteins in the proteome of *A. thaliana* and 8,041 unique proteins in *O. sativa*, which represent 9.7% and 13.9% of the whole proteome respectively.

Previous published estimations of the number of *A. thaliana* unique proteins gave different values ranging from 3,405 to 12,265 proteins (The *Arabidopsis* Genome Initiative, 2000; Mohseni-Zadeh *et al.*, 2004; Wu *et al.*, 2006; Sterck *et al.*, 2007) depending on the protocol used. The smaller value (3,405) comes from the PHYTOPROT project (Mohseni-Zadeh *et al.*, 2004) and were obtained through extensive all-against-all sequence comparisons using the LASSAP software (Glemet and Codani, 1997). The list of unique genes delivered by PHYTOPROT was longer than the list provided by our method but 81% of the unique proteins were shared between both lists. The expertise of additional proteins identified in PHYTOPROT shows that they are members of a PFAM family and, therefore, excluded from our list.

II.C.2. Unique proteins conserved and non-conserved between *Arabidopsis thaliana* and *Oryza sativa*

One protein unique in a given species may have either no, one or several homologs in other species. We named U[1:0] the unique proteins in one species with no homolog in the other one, U[1:1] the unique proteins with only one homolog and U[1:m] the unique proteins with more than one homolog. A 2-letter prefix was added to indicate the plant species when necessary, *i.e.* AtU[1:m] refers to *A. thaliana* unique genes with at least 2 homologs in the *O. sativa* genome. Both U[1:1] and U[1:m] are conserved single copy genes in the reference genome (thereafter called conserved single copy genes) and are respectively qualified as pan-orthologs and syn-orthologs according to Blair *et al.* (Blair *et al.*, 2005).

After sequence comparison based on BLASTp, 995 (3.7% of the whole *A. thaliana* proteome) and 6,418 (11.1% of the whole *O. sativa* proteome) unique genes were classified as AtU[1:0] and OsU[1:0] respectively (Figure 6). Sequence conservation between the Liliopsida and core eudicotyledon members of a pair of proteins is a strong support for the gene prediction of U[1:1] and U[1:m] genes. However, an over-prediction of U[1:0] genes remained possible. Thus, we searched for proofs of transcription for the genes coding for the U[1:0] proteins in both plants. We have found transcript sequences for 544 (out of 995) and 1,462 (out of 6,418) U[1:0] proteins from *A. thaliana* or *O. sativa* respectively. This class of proteins for which the corresponding gene structure was sustained by transcript sequences (ESTs and/or cDNA) was named U[1:0]E (for

Expressed) genes. Similarly, the class of unique proteins without homologs in the other plant species and without cognate ESTs was named U[1:0]NE (for No proof of Expression) genes (Figure 6).

In *A. thaliana*, we further analysed possible over-prediction of 451 AtU[1:0]NE proteins searching for corresponding gene expression in CATMA (Gagnot *et al.*, 2008) and Affymetrix (Zimmermann *et al.*, 2004; Zimmermann *et al.*, 2005) transcriptome resources. Statistical proof of expression was found for 311 additional AtU[1:0]NE genes. All together, these data indicated that most of the predicted AtU[1:0] coding genes were expressed and thus actual genes. It was more difficult to conclude on the accuracy of the number of unique genes for *O. sativa* since there remained a large number of OsU[1:0]NE genes (4,956) with not enough available transcriptome data.

Using the 2,570 *A. thaliana* unique proteins as query in a BLASTp against the 8,041 *O. sativa* unique proteins we found 974 pairs of AtU[1:1] proteins and 960 OsU[1:1] when doing the inverse search. Of these genes, 937 shared pairs remained as U[1:1]protein pairs after crossing both lists. A manual check of U[1:1] protein pairs present in only one list showed that differences were due to gene splitting/fusions that may come from either actual events or from gene prediction errors in one of the two genomes. These processes changed an actual U[1:1] relationship into an apparent U[1:m] relationship.

II.C.3. Topological organization of unique genes

Both *A. thaliana* and *O. sativa* have large regions that are still recognizable as duplicated regions (The *Arabidopsis* Genome Initiative, 2000; Guyot and Keller, 2004). We analyzed AtU[1:0], AtU[1:1] and AtU[1:m] gene distribution in *A. thaliana* non-duplicated regions, which contained 15.7% of the nuclear genome. No significant preferential occurrences of AtU[1:0], AtU[1:1] and AtU[1:m] genes were observed inside the apparently non-duplicated regions, where we observed about 18% of them. Therefore, this result showed that most of the genes are unique not because they belong to a genomic region deleted after whole genome duplication, but because of the non-reciprocal local losses between two paralogous duplicated genomic regions.

We also analysed the distribution of each class of unique genes along *A. thaliana* and *O. sativa* chromosomes using a Chi-square test with a confidence level of 99.5% (critical values of 14.86 and 26.76 respectively). All gene classes were evenly distributed among the 5 chromosomes of *A. thaliana* with a Chi-square of 3.91 for U[1:0], 3.95 for U[1:1] and 0.63 for U[1:m] genes. The *O. sativa* distribution was also even for U[1:0] and U[1:m], chi-square of 23.63 and 25.64

respectively, but unequal (Chi-square of 65.10) on U[1:1] genes. Detailed analysis showed that in *O. sativa* genome there was a higher density of U[1:1] genes in chromosome 2 and 3 and a lower density in chromosome 11 and 12. This particular distribution is unexpected since chromosomes 11 and 12 are the only two rice chromosomes that do not show evidence for large regional duplications with any other rice chromosomes (Guyot and Keller, 2004; Rice Chromosomes 11 and 12 Sequencing Consortia, 2005) [Recent results should be considered here: (Salse *et al.*, 2008)]. The recent duplication described between the first 3 Mb of the chromosomes 11 and 12 (Paterson *et al.*, 2004; Rice Chromosomes 11 and 12 Sequencing Consortia, 2005; Yu *et al.*, 2005) only covers 11% of their size which is not sufficient to explain the low number of unique genes observed within each chromosome (60% of the expected number).

Thus, our results suggest that in *O. sativa*, as well as in *A. thaliana*, non-reciprocal losses between duplicated genomic regions are a frequent mechanism for generating and maintaining unique a set of genes.

II.C.4. Unique gene and protein features

We compared the intron relative numbers, the presence of some TFBS and the protein lengths between random sets of nuclear genes and the 3 groups of unique genes, U[1:0]E, U[1:1] and U[1:m]. All the U[1:0]NE genes and the U[1:0]E genes not fully covered by cognate transcripts were not included in the study due to the uncertainty on their structural annotation (intron number and positions, CDS size). The GC content of all the groups was not significantly dissimilar to the 44.2% in *A. thaliana* and the 53.3% in *O. sativa*.

II.C.4.a. Intron number

This feature separates all the unique genes into two distinct groups. On one side, U[1:0] clustered intron poor genes that had 30% fewer introns than all nuclear genes. On the other side, U[1:m] and U[1:1] genes have a higher number of introns with a density of 1.35 and 1.57 introns per 100 amino acids as compared to 1.09 for all the nuclear genes in *A. thaliana* (Table 2). These differences are the same for rice unique genes. Our results are in agreement with the fact that, in general, evolutionarily conserved genes preferentially accumulate introns (Carmel *et al.*, 2007). Nevertheless, there is no difference in the number of introns in the 5' and 3' UTRs between unique genes and the whole genome. These observations suggest that the pressure of selection that is at work to keep unique a set of orthologous genes in a genome has an effect down to the level of gene

structures mainly in the ORFs. Indeed, functional reasons may be put forward since introns may play a functional role through alternative splicing, effects on gene expression (Brooks *et al.*, 1994; Carmel *et al.*, 2007) or by their involvement in protein transport (Benabdellah *et al.*, 2007).

	All other nuclear genes	U[1:0]E genes	U[1:1] genes	U[1:m] genes
A. thaliana				
Mean intron number	4.28	0.98	5.01	4.33
Mean protein size	392.88	133.53	318.05	318.75
Median protein size	352.00	107.00	262.00	249.50
Mean intron number / 100 aa	1.09	0.73	1.57	1.35
TATA-box presence	18.8 %	26.8 %	10.3 %	11.9 %
TELO-box presence	10.9 %	10.0 %	15.2 %	14.7 %
SORLIP2-box presence	11.9 %	14.1 %	16.1 %	15.2 %
CAAT-box presence	26.2 %	27.7 %	34.9 %	40.3 %
O. sativa				
Mean intron number	3.85	0.85	4.89	4.10
Mean protein size	406.06	142.82	321.15	319.10
Median protein size	362.00	117.00	262.00	266.00
Mean intron number / 100 aa	0.95	0.60	1.52	1.28
TATA-box presence	17.7 %	16.9 %	4.1 %	7.0 %
TELO-box presence	9.1 %	6.4 %	11.0 %	12.9 %
SORLIP2-box presence	38.1 %	31.4 %	41.2 %	36.9 %
CAAT-box presence	34.5 %	33 %	38.3 %	31.2 %

Table 2 – Features of unique genes and their promoter

Only genes with CDS fully covered by transcripts (EST and/or cDNA) were used for the determination of intron numbers and protein sizes. TFBS in promoter regions were searched for only in promoters of genes with a UTR longer than 50 nucleotides as shown by at least one cognate transcript. The complete nuclear gene set minus the 3 classes of unique genes defines the 'All other nuclear genes' class.

II.C.4.b. Transcription factor binding sites (TFBS) in promoter sequences

In the whole genome of *A. thaliana* and *O. sativa* we found respectively 20% and 16% of genes with a TATA-box in their promoters. Comparisons with the frequency of these two well characterized TFBS present in promoters of unique genes split them in two groups: the U[1:0] class on one side and the U[1:m] and U[1:1] classes on the other side. On one hand, the promoters of *Arabidopsis* U[1:m] and U[1:1] genes contains the same relative number of TATA-box (Chi-squared test, P-value=0.40) and they have a significantly lower frequency of TATA-box (Chi-squared test, P-value=2.3e-14) than the other nuclear genes (Table 2). On the other hand, TELO-box presence was significantly higher in AtU[1:m] and AtU[1:1] genes than in the other nuclear

genes (Chi-squared test, P-value=0.0057). The same differences are observed in unique *O. sativa* genes (Table 2). The two other TFBS analysed, SORLIP2 (Hudson and Quail, 2003; Tremousaygue *et al.*, 2003) and CAAT (Bucher and Trifonov, 1988) boxes, present slight variations in each class when compared with whole genome distribution, but these variations were not consistent in both species (Table 2). The different frequencies of TATA and TELO boxes observed in the promoter sequences of unique genes cluster them as the intron density criteria: the class U[1:0] on one side and the two classes U[1:1] and U[1:m] on the other side. This particular clustering conserved in both *A. thaliana* and *O. sativa* is discussed below.

II.C.4.c. Protein length

We compared the size distribution of each group of unique proteins in the two species (Figure 7). On average, unique genes coded for shorter proteins than the whole genome. This is particularly evident for U[1:0] genes in both *A. thaliana* and *O. sativa* showing a mean length and a size distribution of the proteins smaller (Wilcoxon test, P-values < 2.2e-16) than in the other classes of unique genes (Figure 7). Indeed, the median size of the not unique *A. thaliana* proteins is 352 aa while the median value for the U[1:0]E proteins is only 107 aa, *i.e.* about 70% smaller (Table 2). While the displacement of the size distribution of unique genes towards the small values was shown in both plant genomes studied, it was less important in U[1:1] and U[1:m] proteins but still significant (Wilcoxon test, P-values < 1e-14). The size distribution of these two groups of conserved single copy genes had a maximum around 150 aa and is localized between the size distribution of U[1:0] proteins and the size distribution of the whole proteome (Figure 7). We may expect that the number of conserved single copy genes will increase in the near future since more genes coding for short polypeptides will be added to genome annotations. Indeed, the *ab initio* prediction of short ORFs is difficult (Skovgaard *et al.*, 2001; Linial, 2003; Snyder and Gerstein, 2003) and recent results in *A. thaliana* show that a part of the drop in the size distribution of annotated gene products below 100 amino acids (Lease and Walker, 2006) may be due to the rejection by the annotation processes of several small ORFs that turned out to be transcribed and/or under purifying selection (Aubourg *et al.*, 2007; Hanada *et al.*, 2007; Moskal *et al.*, 2007). Similar situations have been reported in mouse, yeast and drosophila where experimental supports and comparative genomics indicate that many short ORFs code for functional elements involved in important biological processes such as cell signalling (Frith *et al.*, 2006; Kastenmayer *et al.*, 2006; Galindo *et al.*, 2007).

In summary, in the *A. thaliana* genome, there are 2,570 unique genes and 995 do not have a homolog in *O. sativa*. Conserved single copy genes are both the 974 *A. thaliana* genes that have only one ortholog and the 601 genes that have more than one homolog in *O. sativa*. In *O. sativa* genome, 8,041 genes are unique and 6,418 do not have a homolog in *A. thaliana*. Furthermore, 960 conserved unique genes have only one ortholog while 663 have more than one ortholog. Even if we might suspect some over-prediction of unique *O. sativa* genes, our results about the common features shared by unique genes are highly similar in both *A. thaliana* and *O. sativa*. First, conserved single copy genes (U[1:1] and U[1:m] classes) have relatively more introns than in the whole genome and their promoter is characterized by a lower presence of TATA-box and a higher presence of TELO-box than in the nuclear genes. Second, unique genes code for shorter proteins than the whole genome and the difference is the highest for unconserved proteins.

II.C.5. Functional features of U[1:0] genes

We recovered the annotated gene functions available for the 544 AtU[1:0]E. Despite the fact that we used *'*annotation*'* in the largest acceptation of the word, only 105 of them have a predicted function (Table 3), *i.e.* 2 to 3 times less than expected from the whole genome (Swarbreck *et al.*, 2008). In the 105 annotated AtU[1:0]E genes we observed 15 genes coding for recognized peptide phytohormones (Farrokhi *et al.*, 2008) including CLAVATA3 and 5 CLAVATA3 related peptides, POLARIS, 3 PROPEP, RALF and N-Hydroxyprolin-rich glycoprotein coding genes. The small peptide phytohormones are involved in signalling roles in defence or non-defence functions (Farrokhi *et al.*, 2008). Most of the peptide phytohormones are proteolytic products of larger propeptides encoded by different genes. Some peptide phytohormones may be clustered based on short motif conservation such as CLAVATA3 group which is characterised by only 12 residues while the remaining parts of the propeptides are highly divergent. When we searched for peptide phytohormones in AtU[1:1] genes, we did not find any even though there were almost 6 times more genes with predicted functions compared to AtU[1:0]E genes.

			With predicted function		Peptide phytohormones		Pro- or Gly-rich proteins	
	Nb	ER (%)	Nb	ER (%)	Nb	ER (%)	Nb	ER (%)
AtU[1:0]E	544	19.7 %	105	27.6 %	15	46.7 %	13	61.5 %
AtU[1:1]	937	6.8 %	610	6.0 %	0	-	4	0.0 %

Table 3 – AtU[1:0]E and AtU[1:1] function comparison

Distribution of AtU[1:0]E and AtU[1:1] genes according their functional annotation and the presence of predicted targeting peptide to the endoplasmatic reticulum (ER).

Another specific feature of the AtU[1:0]E group is to exhibit a relatively high percentage of genes coding for proteins targeted at the endoplasmatic reticulum (Table 3) as pro-peptides coding for secreted peptide phytohormones (Farrokhi *et al.*, 2008). This observation suggests that the AtU[1:0]E group might contain many other not yet characterized genes coding for pro-peptides phytohormones and that might be involved in unknown signalling processes. For instance in the AtU[1:0]E group, we found 13 genes coding for proline or glycine rich-proteins that were mainly predicted to be targeted at the endoplasmic reticulum (Table 3). Additionally, genes encoding for secreted peptides have been reported as having a low intron density (Lease and Walker, 2006) as we observed for the U[1:0] group of genes.

II.D. Orthologs genes analysis

II.D.1. Structural and functional features conserved in At and OsU[1:1] gene pairs

The 937 pairs of U[1:1]genes between *A. thaliana* and *O. sativa* were established on local sequence comparisons (reciprocal best hit or RBH) of U[1:1] gene lists with criteria generally accepted to define an orthology relationship (Tatusov *et al.*, 1997). Nevertheless, to support more strongly the orthology and the functional relationships, we looked for some structural features shared by the two members of U[1:1] pairs.

II.D.1.a. Protein length

Protein lengths of the two members of a U[1:1] pair were highly correlated (Figure 8A) and the slope of the correlation was close to one. Indeed, 456 (49%) out of the 937 pairs had proteins with length differing by less than 5% of the total length. This high conservation in protein length between the proteins of a U[1:1] pair was also illustrated by the fact that in 526 pairs (56%) the difference between the two proteins was less than 20 amino acids. Nevertheless, a small number of U[1:1] pairs were more divergent with, for instance, 77 pairs (8%) showing differences in the protein lengths equal to or higher than 30%. We examined the 24 pairs exhibiting a length difference higher than 200 amino acids, and in 16 cases, the difference could be explained by errors in the predicted gene model of one of the two genes. In 4 out of 16 pairs we found an artifactual fusion or splitting of neighbour genes (Figure 8B) and in 12 out of 16 pairs the difference was due to an erroneous gain or loss of exons (Figure 8C) in one of the two species.

II.D.1.b. Intron position

The conceptual position of introns has been searched in the global alignment of each pair of protein sequences. Nearly 45% of U[1:1] pairs had conserved number and positions of introns, while the mean value for random pairs of conserved unique genes was 0.2% (Table 4). Less stringently, 71% of the U[1:1] pairs exhibited at least one intron at a conserved position as compared to 10.6% in the random pairs. Overall, the high intron conservation is strong evidence for orthology between members of a U[1:1] gene pair, discarding any mechanism of convergence between their sequences. Comparison of gene structures in the U[1:1] pairs also highlights the fact that, since the speciation, the numbers of intron gains or losses are nearly equivalent in the two species. Indeed, the ratio between the number of not conserved introns (in terms of position) in *A. thaliana* and the number of not conserved introns in *O. sativa* is 1.03 (Table 4).

	U[1:1] ortholog pairs	**U[1:1] random pairs**	**Nuclear gene random pairs**
Pairs with all conserved intron positions	44.9 %	0.2 %	0.1 %
Pairs with no conserved intron position	44.4 %	79.6 %	58.9 %
Pairs without any intron	3.7 %	1.0 %	5.1 %
Pairs where only one gene has intron(s)	7.0 %	19.2 %	36 %
Pairs with at least one conserved intron position	71 %	10.6 %	6.1 %
Conserved intron number / total intron number in *A. thaliana*	60.5 %	2.6 %	1.7 %
Conserved intron number / total intron number in *O. sativa*	59.6 %	2.6 %	1.9 %
Number of not conserved introns in *A. thaliana* / not conserved introns in *O. sativa*	1.03	1.02	1.12

Table 4 – Conservation of intron positions in U[1:1] gene pairs

Intron position conservation was tested between 486 U[1:1] gene pairs (pairs in which both genes are supported by full-length transcript), and on random samples of 486 shuffled gene pairs from both species extracted fifty times from U[1:1] genes and from all nuclear genes. Intron position was based on the corresponding protein sequence alignments (ClustalW).

Comparative studies on *A. thaliana* and *O. sativa* genes showed three different evolutionary trends based on the orthology relationships. First, recent duplicated genes are submitted to high loss and gain of introns (Knowles and McLysaght, 2006), second, two orthologous genes tend to keep the same gene structure and only a relatively small number of species-specific introns are observed (Roy and Penny, 2007) and, third, slowly evolving conserved genes are also subject to an elevated rate of intron gain but tend to conserve their introns (Carmel *et al.*, 2007). As a consequence, there

is a negative correlation between density of introns and sequence evolution rate of genes (Carmel *et al.*, 2007). The density and the high conservation of intron positions in conserved unique genes, U[1:1], suggests that these genes are orthologous and slowly evolving genes.

II.D.1.c. Transcription

The methods available to compare the expression of orthologous genes from different species are limited. Since *A. thaliana* and *O. sativa* benefit from large collections of EST and cDNA sequences, we used the number of available cognate transcripts of each member of a U[1:1] pair to estimate and compare their expression levels. In order to avoid sampling bias, we focused our comparison on genes with at least 30 cognate transcripts. Retrieved information showed genes with at least 30 cognate transcripts are in similar proportion in the population of U[1:1] genes as in the whole genome whatever the considered species: 14.6% and 17.3% respectively for *A. thaliana* and 7.2% and 10.1% respectively for *O. sativa*. A correlation (Kendall's test, P-value=1e-6) between the normalized numbers of transcripts in *A. thaliana* and *O. sativa* could be observed for U[1:1] pairs (Figure 9A). We compared this result with the correlation obtained with a random set of gene pairs having a maximum size difference of 20 amino acids to reflect U[1:1] size proximity. The random set contains ten times more gene pairs to compensate for the fact that associating not orthologous genes increases the chance of having at least one gene in the pair with less than 30 ESTs/cDNA. No correlation between the numbers of ESTs within the random set (Kendall's test, P-value=0.26) was found (Figure 9B). Gene expression and evolutionary rate have been shown to be correlated in the genomes of different species (Pal *et al.*, 2001; Krylov *et al.*, 2003; Drummond *et al.*, 2006) including plants (Wright *et al.*, 2004). Our results showed that this correlation held true for the limited set of conserved unique genes in *A. thaliana* and *O. sativa*. Indeed, similarities inside U[1:1] protein pairs coming from highly transcribed genes, *i.e.* with at least 30 cognate transcripts, were higher than similarities in the lowly transcribed U[1:1] pairs (Figure 9C). Therefore, the features expected for genes responsible for the same biological function, *i.e.* conservation both in sequence and in level of transcription as well as the positive correlation between them, are strongly observed between genes in U[1:1] pairs indicating their pan-orthology.

II.D.1.d. TFBS conservation

In the previous section, we showed that conserved unique genes have less frequently a TATA-box and more frequently a TELO-box in their promoters than the other genes. Nevertheless,

the general over-representation of one TFBS in the unique gene promoter set does not mean that TFBS are conserved in the two promoters of pan-orthologs. Therefore, we searched for the number of simultaneous TATA-box or TELO-box presence on both promoters of each U[1:1] gene pair. Surprisingly, the percentage of pan-orthologs that presented a TATA-box motif within both promoters was only 0.8% and is not significantly different (Chi-squared test, P-value=0.13) than the expected value, *i.e.* the value observed in promoters of randomly selected pairs of genes (0.4%). In contrary, the simultaneous presence of a TELO-box motif within both promoters of a U[1:1] pair was significantly higher (Chi-squared test, P-value=5.22e-5) than found in random pairs (3.8% compared to 1.6%). In order to complete the promoter comparison between *A. thaliana* and *O. sativa* pan-orthologs, we used the CONREAL (Berezikov *et al.*, 2005) and CREDO (Hindemitt and Mayer, 2005) packages to find any other conserved motifs, *i.e.* known or not known putative TFBS. This phylogenetic footprinting approach did not highlight a promoter sequence conservation different than that detected in random pairs of promoters. Additionally, the global analysis of all pan-ortholog promoter pairs with Motif sampler (Thijs *et al.*, 2001) failed to discover over-represented motifs excepted the previously identified TELO-box. Thus, contrary to our observation of conserved features in the CDS, we found almost no trace of sequence conservation within the promoters of U[1:1] gene pairs even if our dataset of pan-orthologs might be regarded as the best situation to see common regulatory sequences in *A. thaliana* and *O. sativa* promoters. Nevertheless, promoter pairs of pan-orthologs might share conserved TFBS (not over-represented in the unique gene population) which we cannot distinguish from background noise through the comparison of two sequences.

In summary, conserved genes maintained unique in both *A. thaliana* and *O. sativa* have (i) clearly a common origin as indicated by the conservation of the intron positions and the conservation in their product lengths, (ii) no apparent conservation between their promoters which contrasts with (iii) a conservation in their relative transcription level. Nevertheless, the number of ESTs that may be associated to a gene is a general indication of the level of transcription but it is a mixed measurement that is dependent on both high expression in specific situations and expression in a large range of conditions. Transcriptome data from DNA chips inform better on the breadth of expression. Analyses of large transcriptome data collections have shown that *A. thaliana* genes responding to many stimuli are frequently characterized by the presence of a TATA-box, shorter CDS and fewer introns (Walther *et al.*, 2007). Conversely, *A. thaliana* genes controlled by TELO-box have a narrow stimuli response and tend to be larger and have more introns (Walther *et al.*, 2007). In this context, the conserved single copy genes, which rarely contain a TATA-box and are

relatively short genes containing more introns, might constitute a group of genes quite apart in the whole genome.

II.D.2. Are unique *A. thaliana* and *O. sativa* genes conserved as unique in other plants?

We extended our study to other genomes for which our knowledge was not as complete as for the *A. thaliana* and *O. sativa* ones but, nevertheless, with a relatively complete proteome available. Thus, we systematically searched, with our approach, for unique proteins in the available proteomes of *Ostreococcus lucimarinus* and *Physcomitrella patens* (Palenik *et al.*, 2007; Rensing *et al.*, 2008). The nearly complete proteomes of *Populus trichocarpa* (Tuskan *et al.*, 2006) and *Vitis vinifera* (Jaillon *et al.*, 2007) were not used in our phylogenetic analysis in order no to distort our results by an overrepresentation of the core eudicotyledon branch. Two by two comparisons of the unique proteins from the 4 studied species showed that the number of U[1:1] pairs decreased with the evolutionary distance separating the plants. However, the numbers of the observed U[1:1] pairs were always significantly above the number expected by chance (Figure 10). There are about the same number of U[1:1] pairs, ranging from 477 to 503, between *O. lucimarinus* proteome and any one of the 3 other proteomes whatever their total number of proteins (ranging from 26,541 to 57,915). This result suggests that any of the multicellular plant genomes conserved about 500 of the unique 2,691 genes present in the unicellular *O. lucimarinus* genome. Globally, 53% of the 2,691 *O. lucimarinus* unique genes are also present in all or all but one species and 18% of the *O. lucimarinus* unique genes are also present in *P. patens* but not in the core eudicotyledon and Liliopsida plants used in the comparison. Similar results are obtained for *A. thaliana* unique genes with 48% of the unique genes present in all or all but one species, and 22% of *A. thaliana* unique genes only found in *O. sativa*.

The phylogenetic studies of unique gene conservation from *O. lucimarinus* to *A. thaliana* provided a final list of 192 unique genes, the intersection between the two lists (200 and 209) provided by comparisons going in the two opposite directions (Figure 10). We named as U[1:1:1:1] these genes conserved as unique in the 4 studied species. The 192 U[1:1:1:1] genes constitutes a particular subset (genes maintained as single copy in every studied species) of the 4,177 *A. thaliana* core genes defined as conserved in all plants by Vandepoele and Van de Peer (Vandepoele and Van de Peer, 2005). The expected number of U[1:1:1:1] genes, if we assumed a random conservation between *O. lucimarinus* and *A. thaliana*, was only 8.38 genes (Methods section). The 192 genes present in all the studied species came from a common ancestor about 725-1,150 MYA (Hedges *et*

al., 2004; Zimmer *et al.*, 2007) and have been conserved as unique in all the species despite numerous local and segmental duplications expected to have occurred during this long period of time (Jaillon *et al.*, 2007). In comparison, Zimmer *et al.* (Zimmer *et al.*, 2007) have defined 26 pan-ortholog clusters but they have also considered *Cyanidioschyzon merolae* and *Pinus taeda* data and allowed for the exception that a single species might contain paralogs.

Structural features of U[1:1:1:1] genes showed a mean protein length and exon number similar to features in U[1:1] genes as well as the same tendency towards a low TATA-box and a high TELO-box presence in promoters. These characteristics suggest that unique genes underwent the same kind of selection pressure from the common ancestor to the present organisms. An estimation of this pressure was obtained by calculating the synonymous and non-synonymous substitution rates (dN and dS) with Nei-Gojobori's method (Nei and Gojobori, 1986) included in the Codeml program from the PAML package (Yang, 2007). Each gene within a cluster of U[1:1:1:1] genes was paired and compared to every other gene included in the cluster (Table 5). Additionally, the dN/dS rate was computed for U[1:1] gene pairs. Results showed a high selective pressure against non-synonymous substitutions with a median dN/dS ratio of 0.32 for the 937 U[1:1] genes and from 0.25 to 0.41 for unique genes conserved among the three land plants and with a maximum median of 0.79 for pairs including *O. lucimarinus* (Table 5). In comparison, we observed that the median dN/dS ratio calculated from 7,551 alignments of putative *A. thaliana* - *O. sativa* orthologous proteins (RBH, Methods section) is 0.33. One dN/dS ratio of 1 is usually considered as the limit between a negative or a purifying selection, a drift being equal to 1 and a positive selection being higher than 1 (Anisimova *et al.*, 2001; Nekrutenko *et al.*, 2002). Thus, our results show purifying selection pressure onto conserved unique genes in plants and strongly suggest that most of these genes are actual functional pan-orthologs.

	A. thaliana	*O. sativa*	*P. patens*
O. sativa	0.25		
P. patens	0.41	0.33	
O. lucimarinus	0.79	0.73	0.72

Table 5 – dN/dS rates in plant conserved unique genes

Medians of synonymous and non-synonymous substitution rates (dN/dS) among all pairs in U[1:1:1:1] genes were calculated with Nei-Gojobori's method after ClustalW alignments.

II.D.3. Phylogenetic conservation of unique genes and functional implications

The existence of homologs to U[1:1:1:1] genes in other species was searched by BLASTp against the Uniprot database in order to define the range of conservation in other branches of the tree of life. Our results show that 26% of U[1:1:1:1] genes were specific to plants, 13% were conserved in plants and bacteria, 43% could be found in both plants and metazoa, and 18% were conserved in all plants, bacteria and metazoa phyla. This phylogenetic profile shows that 74% of U[1:1:1:1] genes were highly conserved not only in plants but also in other life phyla. This situation implies an ancient origin of these genes and increases the probability for a critical function promoting their conservation. However, no evidence of shared or similar functions can be found in the fraction of U[1:1:1:1] proteins for which functional information has been inferred from sequence homologies. The fraction of unique conserved genes with a functional annotation, *i.e.* 60%, is the same as in all *A. thaliana* nuclear genes (Swarbreck *et al.*, 2008). In order to get information about function and origin of unique plant genes, we explored the predicted subcellular localization of the proteins according to their phylogenetic profile (Table 6). This work was based on the analysis of the 937 U[1:1] proteins since the 192 U[1:1:1:1] proteins constitute too small a set to obtain statistically robust results. Compared to 20,000 random *A. thaliana* nuclear genes, the unique plant genes having homolog(s) only in bacteria frequently encode plastidial proteins since 49.1% of them have a predicted targeting peptide specific to chloroplasts (Table 6). We observed the same tendency within the 192 U[1:1:1:1] proteins. This significant bias (Chi-squared test, P-value=1e-5) suggests that a large part of the subset of unique conserved plant genes may come from DNA transfer from the chloroplast to the nuclear genome. Horizontal transfer from bacteria to plant genome can also explain a fraction of this gene subset. This gene transfer probably predated the speciation between Liliopsida and core eudicotyledons for the concerned U[1:1] genes and is close to the root of the plant phylum for the group of U[1:1:1:1] genes. Our results suggest that, after their transfer to the nucleus, these genes have been submitted to a strong selection pressure that conserved them as unique. This hypothesis is more parsimonious than many independent gene transfer events in each concerned plant species. In their 26 clusters of pan-orthologs, Zimmer *et al.* (Zimmer *et al.*, 2007) also suggest a DNA transfer from organellar genome, mainly from mitochondria. Our observations on the U[1:1] gene population showed that transfer from mitochondria was also significant (Chi-squared test, P-value=0.0002) but less important than from chloroplasts (Table 6).

	Gene number	Predicted targeting				Promoter	
		Plastid	Mito.	Nucleus	ER	TATA	TELO
Random nuclear genes	**20,000**	**6.0 %**	**3.6 %**	**5.6 %**	**14.7 %**	**19.9 %**	**12.2 %**
plant	50.5 %	4.0 %	3.1 %	6.3 %	15.5 %	22.1 %	9.2 %
plant + bacteria	5.5 %	28.5 %	3.5 %	1.8 %	11.6 %	14.6 %	5.7 %
plant + metazoa	20.4 %	1.9 %	2.4 %	7.7 %	8.0 %	16.5 %	20.5 %
plant + bacteria + metazoa	23.6 %	8.4 %	5.7 %	3.2 %	19.4 %	20.7 %	10.7 %
U[1:1] genes	**937**	**15.5 %**	**6.0 %**	**4.9 %**	**6.8 %**	**10.3 %**	**15.2 %**
plant	49.7 %	16.5 %	5.4 %	5.8 %	8.1 %	8.4 %	16.2 %
plant + bacteria	11.3 %	49.1 %	3.8 %	1.0 %	2.8 %	17.7 %	4.4 %
plant + metazoa	27.9 %	1.2 %	4.6 %	6.1 %	6.5 %	8.2 %	18.5 %
plant + bacteria + metazoa	11.1 %	12. 5 %	14.4 %	1.9 %	5.8 %	18.0 %	11.5 %

Table 6 – Phylogenetic profile, subcellular localization and promoter of U[1:1] genes and proteins

Phylum conservation of 937 U[1:1] genes and 20,000 random nuclear genes was obtained by BLASTp of conserved unique *A. thaliana* proteins against the Uniprot database. Gene number (column 1) shows the relative number of genes with a sequence similarity suggesting homology in different phyla. Subcellular localization (column 2-4) was retrieved from the FLAGdb^{++} database. ER means Endoplasmic Reticulum. Promoter regions of the *A. thaliana* genes were analysed for the presence of TATA (column 5) and TELO (column 6) boxes as previously described in Table 2 and the Methods section.

A second subset of U[1:1] genes with homologs in metazoa (including fungi) must have been conserved from ancient eukaryotic cells through the entire phylum and probably has a critical function. Ancient origin, low divergence rate, presence of TELO-box and dearth of TATA-box (Table 6), suggest that they are, or are related to, housekeeping genes (Tremousaygue *et al.*, 2003; Basehoar *et al.*, 2004) but no evidence could be retrieved from the Gene Ontology annotation due to the high number of unclassified genes. This metazoan conserved subset represents 28% of the 937 U[1:1] genes but, interestingly, this fraction increases to 43% in the 192 U[1:1:1:1] genes.

II.E. Conclusions

We defined 2,570 and 8,041 proteins as unique in *A. thaliana* and *O. sativa* respectively. Unique proteins, products of unique (or single-copy) genes, are proteins with no sequence motif shared by any other protein in the same species. *A. thaliana* unique genes can be further classified according to the number of orthologous genes found in *O. sativa* genome or vice-versa. Final classification included: 451 AtU[1:0]NE, 544 AtU[1:0]E, 974 AtU[1:1], 601 AtU[1:m], 4,956 OsU[1:0]NE, 1,462 OsU[1:0]E, 960 OsU[1:1] and 663 OsU[1:m] genes (Figure 6).

Unique genes are distributed all over the genomes including regions with evidence for segmental duplication and suggesting that unique genes have been created by non-reciprocal local

losses between two paralogous duplicated genomic regions. These non-reciprocal losses may have been directed by a selective pressure according to the structural features present in unique genes conserved in the two species (U[1:1] and U[1:m] genes). These specific features are a relatively small protein size and a high intron density that have been described as evidence of a slow evolution rate (Carmel *et al.*, 2007). From a functional point of view, unique conserved genes are characterized by a rare occurrence of TATA-box and a high occurrence of TELO-box in their promoters suggesting that unique genes could be linked to critical housekeeping functions such as protein catabolism and synthesis, RNA processing or DNA repair (Tremousaygue *et al.*, 2003; Basehoar *et al.*, 2004; Walther *et al.*, 2007). These results differ from previous observations which showed that genes involved in transcription regulation and signal transduction tend to be more duplicated (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004). Additionally, even if unique genes have been conserved in plants, no significant over-representation of TFBS related with photosynthesis or light regulation processes, such as SORLIP2 and CAAT boxes, have been found in *A. thaliana* and *O. sativa* (Table 1).

Unlike conserved single copy genes, the *A. thaliana* and *O. sativa* U[1:0] genes exhibit a low intron density, a normal presence of TFBS in their promoters, and they encode for proteins about 2.5 times shorter when compared to all the nuclear genes. Very short proteins have been reported as proproteins, precursors of regulatory peptides (Lindsey *et al.*, 2002). Despite the fact that the function of 80% of AtU[1:0]E genes and 95% of OsU[1:0]E genes remains unknown, the analysis of the 105 AtU[1:0]E with annotated function seems to reinforce this hypothesis as we have found that many AtU[1:0]E code for known precursors of short peptide phytohormones with signalling roles (Farrokhi *et al.*, 2008).

From a phylogenetic point of view, product length conservation and similar relative transcription level of the 937 pan-orthologous genes in *A. thaliana* and *O. sativa* (U[1:1]) are clear evidence of a common origin. However, intron insertion site conservation is the best proof that couples of U[1.1] have evolved from a common ancestor and are not the consequence of convergence. This intron conservation is also evident for the 192 U[1:1:1:1] genes where dN/dS analysis shows that those genes conserved as unique in very distant photosynthetic species are pan-orthologs under negative selection pressure to keep them in a low divergence rate and unique. This situation reinforces the idea of a probable important conserved function.

It could be suggested that the characterization of pan-orthologs (conserved single copy genes in two or more species) could be noised by the presence of paralogs in the situation where opposite members of a pair of duplicated genes are lost in two daughter species. Nevertheless, our results about conservation of protein sizes, transcription levels and sequence conservation (dN/dS)

argue that, if it is the case, the gene loss occurred before both duplicates diverged enough to allow us to recognized them as paralogs rather than as orthologs.

The phylogenetic profiles of conserved single copy genes and the predicted subcellular location of the corresponding proteins, provides additional information on the origin and the function of these particular genes. An *A. thaliana* subset of unique genes with homologs in plants and bacteria contains 49.1% of genes encoding proteins with targeting peptides specific to the chloroplast. This observation suggests that the origin of this subset of unique genes could be a DNA transfer from chloroplast or bacteria genome posterior to the eukaryote radiation.

Our analysis of the conserved single copy genes, coming in addition to many duplicated gene studies, provides new information on plant gene evolution. Thus, an important part of the genes in only one copy in present plant genomes have an ancient origin and a low divergence rate controlled by a strong selection pressure. The species-specific unique genes that have some structural features in common with the conserved single copy genes are probably recruited from some conserved single copy genes experiencing a rapid divergence linked to a speciation event. However, functions of many of these conserved single copy genes remain unknown. Deeper annotation of small coding sequences that may not be identified by gene finders because of the conservative nature of the prediction algorithms, as well as more experimental data could help to decipher the biological functions of this particular gene population.

CHAPTER III

ANALYSIS OF PROMOTERS OF U[1:1] GENES

III. Analysis of promoters of U[1:1] genes

One of the main aims of this second part of my thesis was the exploitation of a proper set of U[1:1] genes, *i.e.* pairs of probable pan-orthologs, for the detection of new shared transcription factor binding sites (TFBS) through phylogenetic footprinting analyses. Promoter sequence analysis is not trivial since TFBS consist in small (sometimes degenerated) motifs with a low complexity. Moreover, a same TFBS could be found in different positions, comparing to the transcription starting site (TSS), according both to the genes and the species concerned. As a consequence of this diversity, the different programs available to analyse promoter sequences may define different results. Thus, we first tested the different tools and select the most convenient one for a high-throughput analysis that best fits our data. However, the sensitivity and specificity of the available programs are difficult to evaluate due to the lack of a proper benchmark based on promoter sequences with known conserved motifs at defined positions.

Once we have selected the most suitable program for our needs, we used it to analyze the conserved motifs between each U[1:1] pair of promoters (Tagle *et al.*, 1988). Indeed, our set of U[1:1] genes might constitute a valuable set for phylogenetic footprinting as it consists in a list of pairs of probable orthologous genes (see the previous section). For this reason, we hope to take advantage of the promoter sequence of U[1:1] genes to perform pair by pair comparisons in order to highlight common motifs expected to be candidates for TFBS.

Despite from a sequence point of view each pair of U[1:1] genes is independent to the other pairs, unique genes have shown to have some particular characteristics that make them quite different to the rest of nuclear genes. Some of these shared characteristics, as short length, could be explained by the fact that U[1:1] genes belong to a rectricted number of classes of function, as regulatory functions, which could have prompted their reversion to a unique copy after duplication. Such possible analogy of function in U[1:1] genes raises a new question on the possible existence of some other similarities, in particular the possible existence of some 'shared' TFBS between the distinct pairs of U[1:1] promoters. This possibility, which was not previously considered because of the sequence independence of U[1:1] gene pairs, was tested using a protocol that define *ab initio* putative TFBS and their preferential positions to search and compare all the conserved motifs in our U[1:1] promoter set. This approach is expected to provide novel information about the possible use of phylogenetic footprinting as well as new insights in the level of conservation between U[1:1] pairs.

III.A. Detection methods and available software

Many different methods exist for the detection of conserved motifs within promoter sequences. Overall, all the methods can be divided in two major classes depending on the algorithm employed:

String based : String based methods (Tompa, 1999; Jensen and Knudsen, 2000; Vanet *et al.*, 2000), also called word-counting methods, are based on analysis of oligonucleotide frequency (van Helden *et al.*, 1998) or spaced dyads (van Helden *et al.*, 2000). Each pattern representation is compared with the expected number of occurrences and those exceeding a given threshold are considered as overrepresented. Therefore, such threshold can be used to limit the number of false positives. At the end of the algorithm, all the patterns found can be compiled to group the similar ones to find common motifs. Due to their characteristics, string based methods are relatively sensible to background noise and not very prone to the detection of false positives.

Matrix based: Matrix based methods (Lawrence *et al.*, 1993; Bailey and Elkan, 1994; Neuwald *et al.*, 1995; Roth *et al.*, 1998; Thijs *et al.*, 2001) combine a probability matrix and a maximum likelihood estimation algorithm. One of the first matrix based models proposed was dynamic programming algorithm of Needleman and Wunsch (1970) and its local variant (Smith and Waterman, 1981) which uses an identity matrix that gives a score to two aligned data point similarity to find the most optimal alignment. Once the positions are scored, a maximum likelihood estimation is used to distinguish the hidden motif from the background. The most frequent methods to do so are Expectation Maximization (EM) or its stochastic equivalent, Gibbs sampling (Dempster *et al.*, 1977; Geman and Geman, 1984). The combination of matrixes and probabilistic calculations makes matrix based methods not very sensible to background noise but relatively prone to the detection of false positives.

One of the main differences between string and matrix based methods are their sensitivities to background noise. For the analysis of promoter sequences between different species, one can expect that conserved motifs would be hidden in a noisy background which favour matrix based methods over string based methods. From all the different free programs, I have decided to use a combination of matrix based methods which include programs developed for global sequence alignment as well as some methods developed for local alignment which take into consideration some of the promoter characteristics such as the small size of conserved regions or the numerous

non-conserved sequences between them. Some software such as RSAT (van Helden, 2003; Thomas-Chollier *et al.*, 2008) were not considered in our analysis because they are lacking a non web-based version that could be implemented for a high-throughput analysis of the promoters. The selected programs are the following:

- **DIALIGN** : This method was proposed by Morgenstern *et al.* (1998) as an alternative to classic alignment methods. Unlike methods based on Needleman and Wunsch algorithm (Needleman and Wunsch, 1970), Morgenstern's method depends on pairwise and multiple alignment based on whole segments of sequences without gap penalties. To obtain such approach, the DIALIGN algorithm compares diagonals of sequence segments of various lengths. These diagonals, which are the representation of dot-matrix comparison of two sequences, are weighted according to a probabilistic consideration (Morgenstern *et al.*, 1996). Once all diagonals are weighted, the algorithm tries to obtain the set of diagonals with maximum sums of weights: this would be the best alignment possible. The DIALIGN method has some advantages compared to other methods:
    - No gap penalty which allows comparing sequences with local conservation separated by multiple non-conserved sequences.
    - As Needleman and Wunsch algorithm it can be used with different weight matrixes.
    - More adapted to detect recombination and transpositions as it does not try to extend the alignment while minimizing the number of gaps.

- **FootPrinter** : FootPrinter (Blanchette and Tompa, 2003) is an implementation of a previously published algorithm by the same authors (Blanchette *et al.*, 2002). Unlike other programs available, FootPrinter does not try to align multiple sequences to detect regions of conserved similarities. Instead, FootPrinter uses a 'phylogenetic footprinting' approach to discover the conserved motifs. Given a number of orthologous regulatory regions and a phylogenetic tree relating them, the algorithm is capable to detect sequences which have evolved more slowly than the surrounding regions. Main characteristics of FootPrinter algorithm are:
    - Low sensibility to divergence: it works even if sequences are too diverged to be accurately aligned.
    - High sensitivity to detect small conserved regions of the size of many regulatory sequences (5-20 bases) from the background noise.

- About 10 time quicker than DIALIGN on large data sets but similar speed than other methods such CLUSTALW and MEME.
- Does not rely on multiple sequence alignments but needs a correct phylogenetic tree as input.

- **MEME** : MEME (Bailey and Elkan, 1994) is a software implementation of mixture model algorithm, an extension of finite mixture model developed by Aitkin and Rubin (1985). MEME only needs the length of the motifs to detect and the sequences set to work with. From this input, mixture model unsupervised algorithm is capable to discover different motifs with different numbers of occurrences in a single dataset by optimizing the e-value of a statistic related to the level of conservation information of the motif. Some of the features of MEME are:
  - It describes the probability of occurrence of each letter in each position of the motif.
  - It does not work with gaps so does not use gap penalty. However, patterns with variable gaps are split in different motifs.
  - MEME does not consider the position of the motifs so they can be anywhere in sequences.

- **AlignACE** : Described by Hughes *et al.* in 2000, AlignACE is a Gibbs sampling algorithm to identify the over-represented motifs in a given set of DNA sequences. It scores the specificity independently of the method used to find the motifs so it could be used to compare different motif-finding algorithms.

- **MotifSampler** : MotifSampler (Thijs *et al.*, 2001) is, as AlignACE, a Gibbs sampling algorithm implementation. The improvement introduced by MotifSampler is the modelling of the background with a higher order Markov model. Such backgrounds can be either constructed from the input data or be based on independent data. Simplest background models just calculate the frequencies of each nucleotide within the sequence (Liu and Lawrence, 1999) while some other advanced background models try to model the varying GC and AT-content in the different regions of the DNA sequence (McCue *et al.*, 2001). Main advantages of MotifSampler are:
  - Different background models can be used to distinguish motifs from background noise.

- It is robust to background noise, especially when using a 3rd or 4th-order background model.
- Parameter settings can be defined to search for a given number of motifs with a certain length and in a variable range of copy number, which, on the other hand can produce different results each time.

- **LAGAN** : LAGAN, and its multiple alignment version MLAGAN, was proposed in 2003 as an efficient tool for large-scale alignments of genomic DNA (Brudno *et al.*, 2003). The algorithm described constructs a rough global map from local alignments of the two sequences by changing their order, and propose a final alignment from computation of the best alignment within a limited area around this rough global alignment. The LAGAN algorithm features are:
  - Unlike other algorithms, it is suitable to align both: very similar (such human and chimpanzee) and very distant (human and fugu) sequences.
  - It has shown to be accurate in the detection of exons.
  - However, it assumes that orthologous regions have been already identified, which imply the need of an external programs and make difficult its automation.
  - Assumes that sequences contain no rearrangements.

- **MAVID** : MAVID (Bray and Pachter, 2004) is a program based on AVID alignment method (Bray *et al.*, 2003) combined with some pre-processed constraints. These pre-computed points try to incorporate biological information into the alignment and include anchor based on *ab initio* gene predictions and protein alignments, and a pre-computed homology map of the sequences that progressive alignment steps must respect.
Some of the features of MAVID are:
  - Efficient on divergent sequences.
  - Alignment uses a guide tree constructed with an iterative method (does not need to calculate all pairwise alignments) which made it feasible for large number of sequences.
  - It incorporates biological data but there is a cost associated to the pre-computing of data.

- **BLASTZ** : Described by Schwartz *et al.* in their human-mouse alignment work (Schwartz *et al.*, 2003), BLASTZ is an independent implementation of the Gapped

BLAST algorithm (Altschul *et al.*, 1997). BLASTZ method works by extending short near-exact matches without allowing gaps until they exceed a certain threshold. Once this threshold is reached, gaps are allowed within the alignments on the regions where nucleotide content is not extremely biased (low complexity sequences). BLASTZ characteristics include:

- Adjustable parameters and thresholds.
- Optional condition that forces matching regions to occur in same order and orientation in both species.

- **CONREAL** : CONREAL (Berezikov *et al.*, 2005) takes advantage from JASPAR (Sandelin *et al.*, 2004) and TRANSFAC (Matys *et al.*, 2003) databases to construct matrixes of known transcription factor binging sites. These binding sites are subsequently used as anchor points for promoter alignments. CONREAL algorithm advantages are:
  - Designed for promoter analysis.
  - It works with matrixes of usually small motifs and does not rely on sequence comparisons, so it can be used with more divergent species.
  - Uses matrixes of known TFBS which increases its accuracy.
  - However, it can not detect *de novo* motifs.

All the programs presented here, except CLUSTALW, are included in two web based tools that provide an easy to interpret graphical output of the results. These packages are:

- **CREDO PACKAGE** : CREDO is the acronym of Cis-Regulatory Element Detection Online (Hindemitt and Mayer, 2005), a web based tool that integrates, combines and visualizes the analyses of AlignACE, DIALIGN, FootPrinter, MEME and MotifSampler. CREDO is available at http://mips.gsf.de/proj/regulomips/credo.htm
- **CONREAL PACKAGE** : The CONREAL (Berezikov *et al.*, 2005), or CONserved Regulatory Elements anchored ALignment, is a web server that provides an interface that combines CONREAL, LAGAN, BLASTZ and MAVID methods. CONREAL is available at http://conreal.niob.knaw.nl

III.B. Software comparison

As previously explained, the aim of the analysis of ortholog promoters in this work is the detection of conserved motifs that we can consider as putative TFBS. Such approach needs the use

of a high-throughput method combining high sensibility and specificity. We can found in literature several papers that benchmark the different methods available and assess the choice of the best motif detection tool. Such benchmarks can be basically done in two ways using:

- Real data: used by Tompa *et al.* (2005) to assess 13 different motif detection tools, including some of the previously described tools as AlignACE, MEME and MotifSampler, as well as other tools as Consensus, MITRA and Weeder. Performance was tested evaluating detection capacity of known motifs at their real position and orientation within a background model. However, their benchmarking protocol was limited as it did not allow sequence comparisons (*i.e.* phylogenetic footprinting from FootPrinter) as well as it only allowed one motif detection.
- Simulated data: used by Pollard *et al.* in their benchmark (Pollard *et al.*, 2004) and its correction (Pollard *et al.*, 2004). In these tests, sequence evolution was simulated from a random sequence conserving the nucleotide frequencies, and including insertions, deletions and constraining blocks. Such approach profits of ancestral sequence knowledge to assess the detection of conserved motifs. However, results were dependent on simulation assumptions which can introduce biases towards those algorithms more adapted to them.

Overall, the methods tested in each paper are different and nowadays no benchmark paper including all the selected programs has been published. Additionally, both Tompa *et al*. and Pollard *et al.* as well other published results show that selection of the best program is highly dependent of the sequence characteristics such as divergence distance.

In order to decide which of the selected programs best fits with our sequences, we take profit of graphical interface of CREDO and CONREAL to visually compare and expertise the results of a small custom benchmark. Such custom benchmark was performed by comparing the promoter sequences of three pairs of orthologous promoters from our U[1:1] set and three pairs of control sequences.

III.B.1. Benchmark parameters

I randomly selected three pairs of *Arabidopsis thaliana* and *Oryza sativa* U[1:1] genes with some conserved characteristics to construct the testing input. As discussed later, the use of only two sequences reduces the significance of the results. The characteristics criterion selected included the conservation of the predicted function, cellular localization and the same PFAM motif(s). Such

characteristics were consider to select pairs of probable orthologs with all the chances that they have a conserved biological functions, which would have increased probabilities of sharing common regulatory motifs. The selected pairs of U[1:1] genes were:

- OS01G44210-AT1G75350 coding for a 'ribosomal protein L31 protein'
- OS01G71190-AT4G28660 coding for a 'photosystem II reaction centre W protein'
- OS03G60370-AT1G09870 coding for a 'Phosphatase'

All the selected genes have a defined TSS according to the following criteria: 5'UTR of at least 50 bases supported by at least one cognate EST or cDNA sequence. The 1000 nucleotides upstream the defined TSS were retrieved to recover the selected promoter sequences. These selected sequences have high probabilities of sharing TFBS within their promoter sequences but we do not know which conserved motifs they contain and in which position. To minimize this problem and as an additional evaluation criterion, I inserted a large known binding site in both sequences. The selected motif was the GCC box, a 11 bases binding of Ethylene Response Elements (ERE) proteins with the following sequence: TAAGAGCCGCC (Ohme-Takagi and Shinshi, 1995). GCC box was inserted in different positions on each sequence to increase the detection difficulty. The insertion position selected were position -900 in *Oryza sativa* promoters and position -700 in *Arabidopsis thaliana* promoters. Once the GCC box inserted, the pairs of U[1:1] promoters obtained were used as input for the tested programs. On the other hand, a negative control sample was obtained by randomly mixing the original promoter sequences (without GCC box insertion) of the three pairs.

CONREAL package including CONREAL, LAGAN, BLASTZ and MAVID methods was used with default settings. In CREDO package, the 'Phylogenetic Footprinting II' model for distantly related and, therefore, divergent species was selected on presetting options. Global settings were left as default except background model which was changed to plant model.

III.B.2. Benchmark results

CREDO results are presented graphically in a single page showing all the motifs found by the different programs of the package (Figure 11). Each motif found by each program is coloured and a summary view is provided to see the number of programs that have predicted the motif. It is also possible to access to a LOGO view of the detected sequence which is useful when multiple motifs of similar colours are displayed at once. The tests using the CREDO package showed that:

- **DIALIGN** : Search for motifs longer than the inserted motif which prevents it to be detected. However, its single prediction have matched in OS01G44210-AT1G75350 pair with one of CONREAL, LAGAN and MAVID predictions, and in OS03G60370-

AT1G09870 pair with one of the motifs found by FootPrinter, MEME and MotifSampler methods. The non-detection of small motifs could explain why it did not produce any output with the control sample.

- **FootPrinter** : Always find the inserted motif as well as many other motifs almost all along the sequences. However, as most of these extra motifs are not found by any other method, it could be argued that some of them are in fact false positive and not the result of an improved detection method. One way to discard false positives is to reduce the maximum number of mutations but it could results in no output at all even in promoters of orthologous unique genes.
- **MotifSampler** : As FootPrinter, it is a stochastic method that gave better results with multiple iterations (1000 in our test). It found the inserted motifs all the times with less false positives than FootPrinter.
- **MEME** : Always find the inserted motif as well as other motifs. In fact, the results obtained are similar to MotifSampler and matched with some of the FootPrinter predictions.
- **AlignACE** : Can detect motifs smaller than other methods but most of those do not match with any other method predictions. Regarding the inserted motif, AlignACE only partially detects it in the OS01G44210-AT1G75350 couple.

The results of CONREAL web-page combine a graphical output of each method with a sequence alignment view highlighting the conserved motifs found (Figure 12). Unlike CREDO package, the graphic output shows the conserved motifs found without the need to check the sequence but, on the other hand, each method results are separated by sequence alignment. Results obtained with CONREAL can be summarized in two points:

- **CONREAL, LAGAN and MAVID** : They gave all the three quite similar results as already noted by Berezikov *et al.* in their paper (Berezikov *et al.*, 2005). However, only CONREAL algorithm was capable to detect the inserted GCC box.
- **BLASTZ** : Produced no result in two cases and only a short and highly degenerated motif in OS01G71190-AT4G28660 pair. Sequences were aligned but, because no gap was allowed, the extended local alignments did not exceed the minimum threshold to be considered as a motif. It needs the use of less divergent sequences or motifs fully conserved on the same positions to work.

III.B.3. Program selection

The benchmark results split in two groups of programs with similar results and a third group of programs with no or discordant results. On one hand, we find the FootPrinter, MotifSampler and MEME algorithms which have always detected the inserted motif but at cost of other multiple motif detection. Those motifs are possibly false positives according to the fact that these protocols have predicted conserved motifs on control sample. However, it is difficult to assess if these predicted motifs are false positives or not and we can only classify them by comparing other algorithm prediction. On the other hand, CONREAL, LAGAN and MAVID show less predicted motifs (with some of them sharing a similar topological organization) which could be interpreted as a good evidence of their specificity. However, LAGAN and MAVID fail to detect the inserted motif which could indicate that the number of predictions is low due to reduced sensibility. Lastly, we have BLASTZ which produced no output, DIALIGN which only detected long conserved motifs making it too specific, the opposite of AlignACE which highlights plenty of small conserved motifs.

According to these results, which of these programs produces the alignments better adapted to our samples? To give an answer to this question is tricky for various reasons. The more obvious reason is that the sample size is so small that results can not be considered otherwise than an indication. Larger benchmarking tests could provide a more detailed review of the algorithms, however, it is not the aim of this work to compare the different methods further than to obtain a quick guideline of them. Another flaw of the comparison is found on the selection of input sequences. Random promoters with conserved nucleotide frequencies were first considered as input sequences. However, the number of parameters that have to be considered to model real promoter complexity was too high. One possibility to model such complexity was the use of a simulation platform as ROSE (Stoye *et al.*, 1998) or RSAT (Thomas-Chollier *et al.*, 2008) which would imply a new benchmark to evaluate the different evolution simulation algorithms to select the one that best reproduces the promoter evolution. Random promoter sequences were then discarded in favour of the use of real promoters. On these promoters we solve the problem of promoter complexity but like with random sequences, we ignore the sequences and positions of conserved motifs. This lack of knowledge prevents the classification of detected motifs as true or false positives. We have tried to solve this flaw with insertion of the GCC box in precise positions as positive control. However, the detection or not detection of only one long conserved motif cannot be considered as an unequivocal signal of algorithm performance. Moreover, we should not forget that we are working with only two highly divergent sequences that we expect to contain small conserved regions but it is possible that TFBS have changed since speciation (Li *et al.*, 2005). The use of more sequences to construct multiple pair wise or multi alignments could provide more refined detections and increase

relevance of results but, from the species available, when this work has been carried out, only *Arabidopsis thaliana* and *Oryza sativa* have enough transcript sequences to define TSS with accuracy. For all these reasons, graphical output of CREDO and CONREAL packages can give visual guideline of motif confidences based on the number of programs that detect the same motif but no final conclusions about them. A more complete benchmark including several motifs of different sizes and level of conservation inserted in different positions could give more information about software efficiency. Unfortunately, such evaluation was not the aim of the thesis and we lacked time to do it. Nevertheless, we have used the obtained information to assess the selection of appropriate software.

Once programs with no or discordant results have been discarded, the decision is to select an algorithm in each group of programs with similar motif detections. The final software selection was based on another important point for the analysis of thousand of cases: the availability of an offline version, their speed and overall difficulty to integrate them in a high-throughput protocol that tests and evaluates all the obtained results. From this point of view, LAGAN, MAVID and DIALIGN output have the inconvenience to present their results as sequence alignments containing the conserved motifs. Such output can be useful for posterior divergence age calculation based on the alignments but it should be carefully parsed to extract the motifs. FootPrinter implementation produces a motif list output in an html file suitable for computational parsing. MEME output is presented in blocks formats containing all the info making it, comparatively, more difficult to parse. Therefore, MotifSampler is the method that best fulfilled the wanted conditions: it has offline tools with an easy to parse tabulated file result (GFF format), it is fast comparing to the others, and has been able to detect the positive control. Additionally, MotifSampler is a stochastic method that might highlight previously undetected motifs. However, this characteristic can increase the number of false positives in results. MotifSampler results should therefore be carefully scrutinazed.

III.C. Promoter comparisons: Phylogenetic footprinting

Once MotifSampler was defined as the most suitable program for our needs we have run the analysis of promoter sequences from unique genes. However, we have decided to complete the MotifSampler approach with an exhaustive search of all the conserved motifs within each pair of promoters. Indeed, the fact to work with only two sequences allows considering the motif diversity in an exhaustive way. This search combined with different filters to eliminate possible false positives could be therefore compared with a sample of randomly paired promoters in order to

evaluate the possible background noise. With our list of U[1:1] genes, we hoped to be in the best conditions to test phylogenetic footprinting between *Arabidopsis* and rice despite their divergence.

III.C.1. Promoter sets

To assure the maximum accuracy in promoter comparisons only promoter sequences from U[1:1] genes with a well defined TSS (Transcription Starting Site) were used in our analyses. Based on previous work (Bernard *et al.*, 2006), the conditions employed to decide which genes have a well defined TSS are (i) their 5'UTR must be at least 50 bases long, and (ii) the UTR regions are experimentally defined by at least one cognate EST or cDNA sequence (after their spliced alignment on the full genome). The 1000 nucleotides upstream the defined TSS are then retrieved and used as promoter sequences for the following analyses. All the information used for this promoter definition is extracted from the FLAGdb^{++} database (Samson *et al.*, 2004).

With the previous criteria, when the 937 U[1:1] gene pairs are searched for couples that have a defined TSS in both *Arabidopsis thaliana* and *Oryza sativa*, 486 U[1:1] promoters pairs can be selected. The results for the 486 U[1:1] promoter analyses have been compared to the results obtained with a control set made of 4,860 promoter pairs randomly selected among the 13,261 not unique genes in *Arabidopsis thaliana* and the 15,638 not unique genes in *Oryza sativa* with a well-defined TSS. Therefore, the control set is constituted of promoter pairs (one promoter from *Arabidopsis thaliana* and the other one from *Oryza sativa*) coming from genes without sequence and function relationships. In order to reduce the complexity, searching for conserved motifs was focused on the plus strand, following the direction of transcription.

III.C.2. Used methods

III.C.2.a. Development of DECOMO

The first step on sequence analysis consists on an evaluation of the sequence divergence and the number of detected conserved motifs. To do so, we adopted an exhaustive approach to evaluate all the conserved motifs taking in consideration only the nucleotide sequence. A program, named DECOMO (for DEscription of COnserved MOtifs), was wrote *de novo* in PERL to exhaustively compare sequences two by two and retrieve all the conserved motifs longer than a given size with a maximum number of mismatches that can be null (exact match). Once selected these two parameters, the program proceeds as follows (Figure 13):

- The first pair of promoter sequences is retrieved.
- The first nucleotide of sequence 1 is kept as the currently query nucleotide and compared with the first nucleotide of sequence 2, the compared nucleotide. If they match, the second nucleotide of sequence 1 is recovered and compared with the second nucleotide in sequence 2. Motif is elongated this way storing each time the matched nucleotides until a mismatch is found.
    - If the motif kept is longer than the minimum desired size (input parameter), its sequence and starting position on each species are saved before the program continues.
    - When a mismatch occurs, the situation is evaluated according to the selected parameters:
        - If the number of mismatches already found is inferior to the maximum allowed, nucleotides would be substituted by a '-' symbol in stored motif and process goes on.
        - Once the number of mismatches found is equal to the maximum allowed, the elongation process stops and all terminal mismatches ('-') are removed from the stored motif before saving it. Exception is made when motif is composed only by repetitions that are a product of GC microsatellites (Fujimori *et al.*, 2003). In this case, motifs are filtered out to avoid result biases because of counting as conserved these microsatellite repetitions.
- Program recovers the current query nucleotide on sequence 1 and compares it with the second position on sequence 2. This way, nucleotide comparison and motif elongation are tested repeatedly until the end of sequence 2. At this point, the second nucleotide of sequence 1 becomes the query nucleotide and is compared to the first nucleotide of sequence 2. Program goes on looping until all nucleotides of sequence 1 are compared against all nucleotides of sequence 2.
- Once all the nucleotides have been tested the output is formatted in three ways:
    - As the program saves all the motifs longer than the desired size before it tries to elongate the motif, a same core motif can be saved multiple times nested in motifs of increasing length. Additionally, the progression of the comparison along the sequence results in the detection of overlapping motifs with successive starting points. To solve this inconvenience, results can be cleaned by using motif starting point and size to retain only the longest matches and remove the shortest ones. For

instance, if we look for conserved motifs with no mismatches and a minimum length of 8 bases or more, we save the 'AATTGGCCAA' motif and delete the nested motifs, *e.g.* 'AATTGGCCA', 'AATTGGCC', 'ATTGGCCAA', 'ATTGGCCA' and 'TTGGCCAA'.

- Alternatively, it could be interesting to keep only those motifs of a given size to easily compare both species. In the previous case, if 'AATTGGCCAA' motif is found in sequence 1 and 'AATTGGCCTA' motif is found in sequence 2, they will not be detected as equal. In this case, when conserved motif are compared between two sequences, it may be more interesting to eliminate all the motifs longer that a given size and leave only the shared motif 'AATTGGCC.' This has been the option used in the framework of this thesis.
- The maximum mismatch number is set *a priori* without knowledge on the sequence. It is possible that potentially interesting longer conserved motifs with few more mismatches are missed. Optionally, the program can merge the overlapping motifs even if the resulting motifs have more mismatches than set to run the program. For instance, if parameters are set to a minimum length of 6 bases and 1 mismatch, and we compare the sequence 1 'ACTGATACT' with the sequence 2 'AC**A**GAT**C**CT', the program finds 'AC-GAT' and 'GAT-CT' as conserved motifs. By joining these 2 conserved motifs a longer motif 'AC-GAT-CT' may be built up. This option is interesting to find potentially interesting motifs missed at first.

Overall the DECOMO method described here is slow and does not score statistically the motifs found by comparison to a general nucleotide composition. However, it presents the advantage of being accurate and exhaustive, and able to retrieve all the common motifs within two sequences fulfilling the desired criteria, regardless of any consideration other than their sequence conservation. Consequently, if the same motif is found two times in different locations in both species, it is counted twice. As DECOMO retrieves all conserved motifs without scoring them, this program can be the base for other statistical algorithms that evaluate the motifs and decide which are significant and which are not, as well as be useful to evaluate the amount of background noise. For instance, R'MES is a free software package (Hoebeke and Schbath, 2006) to calculate the expected frequency of a given motif through a Markov model. In our case, from an input sequence resulting of the concatenation of all the promoter sequences from each species, R'MES defines two

groups of motifs depending if their frequencies are higher or lower than expected. This result can be crossed with the list computed by DECOMO in order to classify the conserved motifs.

The notion of synteny conservation is one of the filters that can be applied to the results obtained from DECOMO in order to take in consideration the motif order conservation between the two compared sequences. To calculate the maximum number of ordered motifs that can be found in two given sequences, the most exhaustive analysis would consist in a matrix that consider all the possible combinations to recover the largest ones. However, the number of combinations can greatly increase for each additional motif added in the sequences. Because the main goal of this filter is to extract functional motifs (putative TFBS) from the background noise, we can focus our search on the maximum number of ordered motifs. This approach takes advantage of PERL optimisation to sort tables and to minimize the calculations and process time. The 'synteny' filter needs, as input, the file containing a list of all the conserved motifs found within a pair of sequences and their positions in each promoter (DECOMO results) to proceed as follows (Figure 14):

- The start position of each motif on sequence 1 and sequence 2 are retrieved and extra zeroes are added on the left side if they have less than three numbers. Because recovered promoter length is 1000 nucleotides and minimum motif size could be (even useless) one nucleotide, all the motifs positions would be then comprised between 001 and 999, being 1000 the TSS position (step 1). Optionally, motifs can be previously filtered to eliminate GC microsatellites or too degenerated motifs.
- Positions on both species are merged resulting in a list of six digits values followed by a conserved motif (step 2).
- This list is then sorted by ascending order resulting in a list of motifs ordered by its position in the promoter of the first species as well as, for motifs that have the same position, by its position in the promoter of the second species (step 3).
- We compared each motif starting position on both species, after adding motif length to avoid possible overlaps, with the starting positions on each species of those motifs below in the list. We counted all the motifs with start positions in both species upstream of the current one and added the final number in the list before motif start positions (step 4).
- The motif list is then sorted in descending order and the first value of the list is split up and saved as the first ordered element of 'maximum suntan' (step 5).
- The number of motifs starting after the split motif is removed leaving a novel list of merged positions and motifs (step 6).

- Steps 3, 4, 5 and 6 are repeated cyclically until no more motifs are left (step 7). That means one complete loop per motif present in the final list of the maximum number of ordered motifs (step 8).

The list of motifs saved in this protocol (Figure 14, step 8) contains the maximum number of motifs in a conserved order found in two sequences (in the two species for our purpose). It should be noted that this algorithm is useful to count the 'maximum synteny' but motif sequences and positions listed correspond to only one of the (maybe) many possible combinations. Lastly, to control the proper working of this DECOMO filter, I took a sample of sequence comparisons with a variable number of conserved motifs and checked by hand the 'maximum synteny' using TOUCAN software (Aerts *et al.*, 2003) to visualize it. In all the tested examples, even on the more complex cases, the maximum number of ordered motifs delivered by the synteny filter was confirmed and no better alternative could be found. The proposed protocol proved, then, its validity to find maximum motif synteny between two sequences while remaining quick and easy to integrate it in a high-throughput workflow.

III.C.2.b. MotifSampler analysis

As previously explained, MotifSampler (Thijs *et al.*, 2001) is not a possible filter for an exhaustive analysis but a stochastic detection method *per se*. Therefore, and for comparison purposes, the same 486 pairs of promoter sequences from U[1:1] genes and the random sample of 4,860 promoter pairs used with DECOMO were analyzed using MotifSampler program. To take profit of stochastic characteristics the input parameters were the following:

- Program was set up to run 20 times per promoter pair to search for up to ten motifs,
- Prior probability was set to 0.9 to allow a large motif degeneracy,
- Due to inclusion of degeneracy, which was not present in DECOMO analysis, motif search was reduced to motifs of 8 nucleotides only,
- Plant background model available in the MotifSampler webpage was used.

Once the 20 runs for each promoter pair were done, the obtained motifs were gathered and filtered by an external PERL script to eliminate possible overlapping motifs.

III.C.3. Results

III.C.3.a. Number of conserved motifs

As explained in the introduction, U[1:1] ortholog promoters are expected to share conserved motifs. However, we ignore how many and if the motifs are specific to conserved unique genes or are conserved by chance in the promoters. To try to answer this question, we have used the previously explained DECOMO method to evaluate the number of conserved motifs within the promoters of the 486 U[1:1] gene pairs compared with a sample of 4,860 randomly paired promoters. In both samples, the number of not overlapping motifs of length 4, 6, 8, 10 and 12 nucleotides with perfect match were recovered and counted.

In all the cases, results showed no difference between the actual U[1:1] promoter pairs and randomly mixed ones (Figure 15). The similar distribution of different motifs found in both samples suggests that since their speciation around 150 MYA (Wolfe *et al.*, 1989; Chaw *et al.*, 2004), *Arabidopsis thaliana* and *Oryza sativa* promoter sequences have diverged enough to make orthologous U[1:1] promoters undistinguishable from non-orthologous promoters. Such divergence would prevent therefore the distinction of the motifs conserved because of a functional role from those that are due to random similarities.

Many motifs that are putative TFBS have been shown to be located at specific distance from the TSS (Bernard *et al.*, 2006; Yamamoto *et al.*, 2007). Therefore, the conservation of the position of motifs between the two U[1:1] promoters was a first constraint that I tried to exploit in order to discriminate probable functional motifs among random sequence conservation. The motif positions were compared to test the number of motifs found in the first species 50, 100 or 200 nucleotides upstream or downstream to the position in the second species. As expected, the positional constraint reduced the number of motifs shared by both samples. However, whatever the threshold distance used, the number of conserved motifs was similar in both U[1:1] promoters and random promoters (Figure 15). Therefore, if we want to distinguish informative from not informative motifs, is necessary to adopt other filtering strategies.

III.C.3.b. Maximum number of ordered motifs

Diverse promoter analyses have shown that the topological organisation of TFBS is important to regulate and initiate the transcription (Kielbasa *et al.*, 2001; Berendzen *et al.*, 2006; Bellora *et al.*, 2007). Filtering the motifs according to their order along the promoter can thus be useful to distinguish between functional and non functional motifs. The ‘synteny’ filter previously

described has been applied to DECOMO results on motifs of size 6, 8 and 10 nucleotides without mismatch, regardless if they were repeated or not. Unfortunately, obtained results were not as conclusive as expected since, as observed for the number of motifs, the maximum number of ordered motifs that can be found in putative orthologous and non-orthologous promoters are very similar regardless of the motif size (Figure 16). Consequently, many of the detected conserved motifs are probably due to random sequence similarities and are not actual TFBS. Nevertheless, the number of 6 and 8 bases motifs that we found is in the range of the numbers of TFBS observed in well-characterized eukaryotic promoters *i.e.* between 5 and 50 (Arnone and Davidson, 1997; Wilkins, 2002).

As explained when searching for the number of conserved motifs, we expected that some functional TFBS could have conserved a similar position in both species promoters. We thus combined this with the 'synteny' filter. We searched for maximum motif ordering in only considering motifs found 50, 100 and 200 nucleotides upstream or downstream on the second species from their position in the first species. In this case only non-degenerated motifs of size 6 and 8 were analysed to avoid uninformative smaller and bigger motifs (Figure 17 and Table 7 below). Unfortunately, discrimination between U[1:1] and random promoter stays unsuccessful and the distribution curves in both promoter samplers were affected alike by the applied filters.

	DECOMO results								
	Number of motifs	+ position constraint			Max. ordered motifs	+ position constraint			R'MES	
		200 bases	100 bases	50 bases		200 bases	100 bases	50 bases	under-represented	over-represented
Average										
U[1:1]	15.23	6.75	3.90	2.11	5.87	4.89	3.43	2.04	1.14	2.45
Random	16.45	7.33	4.18	2.26	6.14	5.16	3.63	2.15	1.34	2.79
Median										
U[1:1]	15	6	4	2	6	5	3	2	1	2
Random	16	7	4	2	6	5	3	2	1	2

Table 7 – Effect of different filters on DECOMO results for motif of size 8

In a last step, we have tried to use R'MES to classify DECOMO results into two groups of motifs according to their unexpected high or low frequencies. However, the advantages to use one or the other group remain uncertain. First possible option is retaining only motifs present in a higher frequency as generally done in the study of coding regions (Hoebeke and Schbath, 2006). This hypothesis is based on the fact that biological important regions are over-represented within the coding regions while those with negative effect tend to be avoided. While many regulatory motifs are over-represented within promoter sequences because of the fact that few transcription factors can regulate a large number of genes (Wang and Zhang, 2006), it is possible that more specialised

transcription factors present lower frequencies than expected. It could be therefore interesting, in a second option, to retain only motifs present in a lower frequency than expected to analyse them further for putative functional motifs. Nevertheless, the results obtained with both options are similar for U[1:1] and non-orthologous promoters (Table 7). In conclusion, we failed to distinguish candidate TFBS from surrounding randomly identical sequences.

III.C.3.c. MotifSampler results

The DECOMO analysis and the different filters tested provided a wide image of the actual number of conserved motifs but were not able to highlight a difference of complexity between pairs of U[1:1] promoters and a random control set. With the stochastic method MotifSampler (Thijs *et al.*, 2001), we expected that the consideration of not perfect matches and the use of Hidden Markov Model (based on a specific plant background model) to extend Gibbs sampling would allow to go deeper in the promoter analysis.

Here again, all the obtained results show that random and non-random U[1:1] promoters present similar numbers of conserved motifs (Figure 18). The increased sensibility through the motif degeneracy used with MotifSampler is not able to differentiate the two groups of promoter pairs.

III.C.3.d. Positive control

The different analyses done on pairs of U[1:1] promoters suggest that they have diverged too much to show any difference with random promoters. This negative result could be interpreted as a problem in our pipeline. Therefore, to test the proper working of DECOMO and associated filters, a positive control analysis was performed.

The data used for this control are pairs of *Arabidopsis thaliana* and *Raphanus sativus* orthologous genes. The sequence of a 128 kb BAC from radish (Desloire *et al.*, 2003) has been used as query with the FiToCoGene software (http://urgv.evry.inra.fr/projects/FiToCoGene/) to search for a corresponding region in the *Arabidopsis thaliana* genome. A region of the chromosome 1 sharing 70% of genes with the radish sequence has been identified as syntenic, suggesting orthologous relationships between the concerned genes. Among them, 4 pairs have been selected because their orthologous relationships are not noised by tandem duplications in one of the two species (Figure 19). The pairs AT1G63650-RS0110 and AT1G63770-RS0230 encode a transcription factor and an amino-peptidase respectively (deduced from homologies), the two others

encode proteins of unknown function. Promoter sequences (1000 nucleotides) are retrieved from ATG because TSS of these genes are not systematically defined by full length cognate transcripts. The 4 pairs of orthologous promoters have been compared to a random sample comprising the 12 possible combinations between each species promoter. The divergence between *Arabidopsis* and *Raphanus* is estimated to only 16-21 MYA (Koch *et al.*, 2001).

Control promoters as well as random sample were analysed with DECOMO method to retrieve all the conserved motifs of size 6, 8 and 10 bases with no mismatches. Motifs of size 4 bases have not been considered as their small size would 'saturate' the results. On the other hand, motifs of 12 nucleotides have been not considered because of previous results that show that perfectly conserved motifs of such size are improbable. In all the cases, position constraint effect has been additionally tested by setting maximum position difference in species one at 50, 100 and 200 bases upstream or downstream of position in species two. Furthermore, as previously done with U[1:1] promoter pairs, the maximum number of ordered motifs ('maximum synteny' filter) has been searched for. For comparison purpose, the 4 pairs of *Arabidopsis thaliana – Raphanus sativus* orthologous promoters and the 12 random pairs have also been analysed using MotifSampler with the parameters previously described (example in Figure 20). All the results obtained with the motifs of 8 nucleotides are presented in the Table 8.

	DECOMO results							**MotifSampler**	
	Number of motifs	+ position constraint			Max. ordered motifs	+ position constraint			Number of motifs	Max. ordered motifs
		200 bases	100 bases	50 bases		200 bases	100 bases	50 bases		
Average										
Control	37	26.5	23.5	15.5	21.25	21.5	20.5	14	61.5	23.75
Random	26.33	13.5	7.58	3.92	8.83	8.67	6.5	3.83	32.42	9.83
Median										
Control	34	24	20	13	18	18.5	16.5	11.5	62.5	23.5
Random	26.5	13.5	8	4	9.5	8.5	6.5	3.5	28.5	8

Table 8 – Number of motifs of 8 bases detected with different methods and filters

Number of conserved motifs of 8 bases detected by DECOMO and MotifSampler for the 4 *Arabidopsis thaliana – Raphanus sativus* promoter pairs used as positive control, and compared to 12 non-orthologous promoter pairs (Random).

The obtained results, with DECOMO as well as MotifSampler, show clear differences in the number of conserved motifs between the orthologous and non-orthologous promoters. These differences are more important after the application of position constraint and the motif order consideration. This situation confirms our previous assumption about the biological importance of the relative position conservation of motifs and the motif ordering. Nevertheless, it should be

considered that this positive control come from very close species and, consequently, many of these motifs may be not functional but conserved through a small divergence time.

If *Arabidopsis thaliana* is perhaps too far from *Oryza sativa* and too close to *Raphanus sativus* in order to perform an efficient footprinting approach, the recent availability of new sequenced genomes opens the possibility to introduce species with intermediate radiation time in our promoter comparisons.

III.C.3.e. Use of other species

During this thesis period, the publications of the nearly whole complete sequences of the *Populus trichocarpa* (Tuskan *et al.*, 2006) and *Vitis vinifera* (Jaillon *et al.*, 2007) genomes allow us to extend our set of orthologous promoters of unique genes. Therefore, we took profit of the well defined pairs of unique orthologs between *Arabidopsis thaliana*, *Oryza sativa, Vitis vinifera* and *Populus trichocarpa* obtained with the protocol defined for the study of the phylogenetic conservation of unique genes (see Chapter II). Neither *Vitis vinifera* nor *Populus trichocarpa* were used in our previous phylogenetic analysis to not over-represent the core eudicotyledon branch. However, in this study we were interested in the analysis of motif conservation between relatively close species (Figure 21). The set of 306 U[1:1:1:1] genes, conserved in single-copy in the 4 considered species, were analysed in the following couples: *Arabidopsis thaliana-Oryza sativa*, *Arabidopsis thaliana-Vitis vinifera*, *Arabidopsis thaliana-Populus trichocarpa* and *Populus trichocarpa-Vitis vinifera.* Results were compared with a random set of 25,000 gene pairs obtained from randomly pairing U[1:1:1:1] genes of each species. In all the cases, promoters were the 1000 bases sequence upstream the ATG due to the lack of information to define the TSS in most of the newly sequenced genomes.

First, the results with DECOMO, with and without 'maximum synteny' filter, showed an apparent correlation between the divergence date and the conservation level (Figure 22, Figure 23). Indeed, we observed in ancient radiations such *Arabidopsis-Oryza* which occurred ~150 MYA (Wolfe *et al.*, 1989; Chaw *et al.*, 2004) a median of 15 conserved motifs while in more recent radiation as *Populus-Vitis* there was a median of 26 conserved motifs. On the other hand, *Vitis vinifera* and *Populus trichocarpa* who diverged nearly at the same moment from *Arabidopsis thaliana* (Jaillon *et al.*, 2007), presented almost identical distribution curves. However, for all the studied plant pairs, the distribution of conserved motifs within random promoter pairs was similar to actual U[1:1] promoter pairs.

These results suggest that the decreasing number of conserved motifs in function of the increasing divergence time is not explained by conservation in orthologous promoters but to a more global sequence similarity in promoters of unique genes (see Chapter II). To clarify this point, a new random sample was made by randomly pairing 3,000 promoters selected within all the nuclear genes of each species and not only within the unique genes (Figure 22, Figure 23). The results are identical to the previous ones except for the pairs of *Populus trichocarpa-Vitis vinifera* for which the expected shift between orthologous and random promoters is observed. Therefore, the promoters of unique genes in Poplar and Grapevine are closer between them, than the other nuclear genes. This relative similarity in promoters of U[1:1:1:1] genes can be explained by a slower divergence rate of conserved unique genes, including their promoter than of the rest of the genome. The analyses of the coding regions of conserved unique genes suggested the same conclusion (see Chapter II). For the other couples of species, the time since their radiation is probably too long to allow us the same observation.

With the introduction of more species, we also have tried to exploit the ability of MotifSampler to work with more than 2 sequences in order to characterize conserved motifs. We expected that with several promoter sequences, the number of conserved non-functional motifs (*i.e.* background noise) will be very low and, therefore, we hope to highlight putative TFBS in actual orthologous promoters. To test this hypothesis, we used MotifSampler with previous parameters to analyse the 306 quadruplets of U[1:1:1:1] promoters from *Arabidopsis thaliana*, *Oryza sativa, Vitis vinifera* and *Populus trichocarpa* and, for comparison, 5,000 random quadruplets of unique promoters (one by species). No shift between the two samples is discernible (Figure 24). At this level, the consideration of the motif ordering was not possible to analyse with our filter since it has been implemented for only two sequences.

III.D. Putative transcription factor binding sites in U[1:1] genes

Characterization of unique genes has been based on the absence of sequence similarity with any other known coding sequence which means that they are therefore independent of each other. However, as we have seen in unique gene analyses (Chapter II), unique genes constitute a particular group within nuclear genes with several specific characteristics. Some of these characteristics, such as a possible role of U[1:0]E secreted peptide phytohormones, introduced the possibility that, despite their sequence independence, the unique genes might be somehow related in their biological

functions. One consequence of this hypothetical relationship would be not only the abnormal TATA- and TELO-box presence observed within their promoter regions, but also the existence of other shared particularities at transcription regulatory level. For this reason, we decided to analyse all U[1:1] promoters in order to find specific putative transcription factor binding sites exhibiting different frequencies in U[1:1] and in other nuclear genes.

The main limit to explore this possibility is that we need a deep knowledge of both the possible motifs to search and their preferential positions. However, despite it exists different databases which include motif sequences and their localization in the promoter regions (Matys *et al.*, 2003; Sandelin *et al.*, 2004), they mainly include data from known motifs obtained from experimental data. Additionally, despite some motif positions are widely conserved in different species, it should also be considered that, in each species, a particular motif might present shift in its functional positions (Shinshi *et al.*, 1995; Buttner and Singh, 1997; Yamamoto *et al.*, 2007). To overcome these problems and include, for each species, as many motif positions as possible, we have used an *ab initio* approach based on preferential position of motifs in promoters to identify valid TFBS candidates. This method, developed in the URGV bioinformatics group by Virginie Bernard (Bernard *et al.*, 2006) is similar to another method based on position conservation for TFBS identification (Yamamoto *et al.*, 2007). The putative transcription factor binding sites predicted with this *ab initio* approach were used to calculate the frequencies of the different putative TFBS within the promoters of unique genes.

III.D.1. *Ab initio* method for the definition of putative TFBS and their preferential positions

Several regulatory elements, such as the TATA-box or the CAAT-box, are preferentially located relative to the TSS. Based on this idea of preferential position, V. Bernard is developing in the framework of her thesis an *ab initio* approach for the identification of motifs within promoter sequences (Bernard *et al.*, 2006). In a first step, this method extracts all occurrences for each motif (minimum and maximum sizes are parameters) in a sample of promoter sequences aligned from TSS. Then, motif distributions are built by using a sliding window of one base long with a one base shift. For each window, the number of promoters containing the motif is given rather than the number of occurrences to avoid favoring repeated motifs. Then, the promoter sequences are divided into two regions. First, the region comprised between -1000 bases and -300 bases relative to the TSS, is used to learn the distribution model using a simple linear regression and to build 99% confidence intervals. Second, the region ranging from -300 bases and TSS is searched for not

evenly distributed motifs according to the distribution model learned, *i.e.* motifs showing a distribution exhibiting a peak. In the cases where a second peak is detected in addition to the first one, the window size is increased step-by-step from 1 base to 100 bases. The method output is a list of preferentially located motifs (PLM) with the preferential position, *i.e.* the position of the peak in the distribution, the peak bounds (thus the peak width) and a Score Maximal Square (SMS) relative to the base line. Such score is the ratio (peak height - base line) / (upper bound - base line) and could be used to filter out those motifs with peaks less significant than a given value (Figure 25).

III.D.2. Definition of overrepresented putative TFBS

III.D.2.a. Criteria definition

The abnormal presence of TATA- and TELO-box within the promoters of conserved unique genes suggests the possible existence of other abnormally frequent TFBS in U[1:1] genes. In order to evaluate this possibility, a preliminary analysis was performed using a first version of *ab initio* detection method in order to evaluate the possible existence of overrepresented TFBS. Main differences with the previously described protocol were the inclusion of fixed peak window of 20 bases instead of an adaptable window, as well as some simplifications in the definition of the confidence interval. Those simplifications were expected to allow a much quicker detection of putative TFBS while remaining significant enough to evaluate the possible abnormal frequencies of TFBS within promoters of unique genes. With the aim to avoid the consideration of motifs not involved in transcription regulation, we focus our analysis on motifs present in preferential positions (*i.e.* position at which a motif has a high probability to be functional) in both *Arabidopsis thaliana* and *Oryza sativa*. This analysis considers the position conservation simultaneously in the two species. Results obtained would be the base for more detailed analyses of promoter sequences.

Input files for *ab initio* detection program included the 14,689 and 17,720 promoter sequences for which a TSS could be defined in *Arabidopsis thaliana* and *Oryza sativa* respectively. The criteria used to define TSS is, as before, a 5'UTR size of at least 50 bases supported by at least one associated EST or cDNA. The 1000 nucleotides upstream the defined TSS were then retrieved and used as promoter sequences input. Once done, promoters were scanned for motifs of sizes ranging from 5 to 8 nucleotides and a SMS higher than 3 (Figure 26).

The *ab initio* method defined a total of 995 motifs in *Arabidopsis thaliana* and 3,606 motifs in *Oryza sativa* with their preferential position. Interestingly those results showed the presence of more than 3 times more motifs with preferential peak position within *Oryza sativa* than *Arabidopsis*

thaliana. Despite similar correlations between the number of sequences analysed and the number of motifs found have been previously observed with the *ab initio* method, differences were not as high as observed in this case. Nevertheless, the actual impact of the number of analyzed sequences on the final results has not been evaluated and, consequently, the possible existence and causes of a bias on the number of regulatory motifs within each analysed species has not been studied.

Once the motifs were defined in each species, we compared them in order to define possible shared motifs. While many of the detected motifs were specific of each species, the comparison highlighted 482 motifs common to both species (Figure 26). The fact that these 482 motifs have been defined as PLMs (preferentially located motifs) in the two species by the *ab initio* method is a further indication of their potential functionality.

Once the sequence and preferential positions of each of the 482 common motifs between *Arabidopsis thaliana* and *Oryza sativa* were defined, we analyzed the possible abnormal presence of PLMs in U[1:1] genes. To do so, we calculated how frequently each PLMs was observed in both promoters of a U[1:1] pair. However, the percentage of U[1:1] pairs sharing a given motif can not be used to detect the possible alterations of PLMs frequencies if not compared with a random sample. Even in this case, they would only represent (despite being the real one) a single observation of motif frequencies in U[1:1] genes. To avoid this limitation, two different random samples were constructed: one from randomly pairing one U[1:1] promoter of each species, and another one from randomly pairing one nuclear gene promoter of each species. The size of both random samples was fixed to 486 randomly paired promoters because it corresponds to the number of actual U[1:1] pairs for which a proper TSS has been defined in both species (Figure 26).

To calculate criteria able to define a possible PLM overrepresentation, a kind of bootstrap analysis was performed: The random sample construction was repeated 100 times and we have searched, each time, which motifs are found on both promoter sequences at least in one of the 486 pairs of promoters. Once all the repetitions were performed, the lists of motifs found were compared to highlight motifs found at least once in each of the 100 runs. This approach provides an idea of which motifs are more likely to be shared between random promoter pairs because this systematic appearance can be considered as a sign of overrepresentation. Interestingly, the results highlighted the presence of 24 motifs found in all the random samples built from U[1:1] promoters (Figure 26). This conservation, which represents nearly one third of the 94 motifs found on average in each run, is almost the double of the 13 motifs found on all the random sample built from nuclear genes. Such difference is the first evidence of a possible overrepresentation of some PLMS within U[1:1] promoters.

This 'bootstrap' like method presents the disadvantage that a motif found in all but one of the 100 runs would not be detected. To solve this problem, it was necessary to find a criteria to define overrepresented PLMS which did not depend on analysis of different repetitions. For this reason, we have decided to calculate the motif frequencies in the 48,600 random pairs (486 random pairs x 100 runs) of each sample and compared them with the list of motifs found for each sample. The analysis of the results for the 24 motifs in U[1:1] random promoters and the 13 motifs in nuclear random promoters defined a minimum presence in both species of at least 400 out of the 48,600 random pairs for the overrepresented motifs. This distribution allows us to set a cut-off frequency of 0.82% as criteria for further analyses of overrepresented putative PLMs.

III.D.2.b. First results: overrepresented TFBS in U[1:1] and nuclear genes

To quantify the number of motifs not previously detected with bootstrap analysis but overrepresented according with our defined criteria, we repeated the search of overrepresented putative PLMs in U[1:1] and nuclear gene random sets based only on PLMs frequencies. This criteria allows us to detect the overrepresentation of 28 motifs in random U[1:1] genes and 17 motifs in random nuclear genes (Table 9). The different numbers of motifs found in the two samples was the first evidence of possible overrepresentation of some PLMs within U[1:1] genes. Such overrepresentation suggests that the different U[1:1] genes could share similar transcription factor binding sites even if they are not homologous. In order to calculate the number of overlapping motifs found in both samples and, therefore, the number of PLMs only overrepresented in U[1:1] promoters, both list of motifs were crossed. The comparison shows that 13 motifs found in all nuclear genes are also found in U[1:1] genes while 15 motifs are specific to U[1:1] genes, and 4 motifs are only found in all nuclear genes (Figure 27, Table 9). As expected, the four non-overlapping motifs unique to nuclear gene random set are related to the TATA-box consensus motifs for which the low presence in U[1:1] genes was previously highlighted during the study of unique genes (Chapter II). The higher frequency of TELO-box could not be confirmed because TELO-box preferential position include the 5'UTR which has not been included in our analyses.

In addition to the absence of TATA-box in U[1:1] promoters, the overrepresented motifs found in the different samples show an interesting characteristic about their similarities. We can observe how most of the putative PLMs could be divided in two subgroups: one formed by a minimal or 'core' sequence, and another which included different 'derived' sequences of those cores (Figure 27, coloured motifs). Indeed, 5 out of the 32 motifs found could be classified as 'core'

sequences and be related with 17 'derived' sequences. This situation was particularly interesting because 5 of the 15 overrepresented motifs found in U[1:1] promoters present no similarities with the rest of the conserved motifs and constitute a class by their own. This classification implies, that in addition to some 'derived' motifs, U[1:1] promoters present a unique set of overrepresented putative TFBS.

Motifs found overrepresented in the random sample of…	**Motif sequence**	**Preferential peak position in *A. thaliana***		**Preferential peak position in *O. sativa***		**Frequency of the motif** *	**Frequency of the motif** *
		Start	**End**	**Start**	**End**	**(%)**	**(%)**
All nuclear genes	TATAA	-40	-21	-40	-21	2.43	0.44
	ATATA	-40	-21	-40	-21	1.70	0.38
	TATATA	-40	-21	-40	-21	0.87	0.09
	TATAAA	-40	-21	-40	-21	1.10	0.12
All nuclear genes AND U[1:1] genes	ACCGG	-200	-1	-320	-41	1.43	2.70
	CGTGG	-220	-41	-280	-1	1.65	2.46
	TGGCC	-180	-41	-240	-1	1.09	2.10
	ATAAA	-40	-21	-40	-21	2.90	0.94
	ACCCG	-160	-1	-280	-1	0.89	1.21
	AGCCCA	-240	-21	-260	-1	1.25	3.48
	ATGGG	-220	-1	-200	-1	2.40	5.17
	GGGCC	-240	-41	-260	-1	3.41	9.14
	TGGGCC	-240	-41	-260	-1	1.86	5.90
	TGGGC	-240	-21	-300	-1	4.92	12.36
	AGGCC	-240	-21	-240	-1	2.40	5.62
	GGACC	-180	-41	-260	-41	0.86	0.99
	GACCC	-200	-1	-200	-41	1.05	0.99
U[1:1]	TTGGGCC	-240	-41	-260	-21	0.20	1.00
	AGGCCCA	-240	-21	-240	-21	0.43	1.56
	AAGCCCA	-240	-41	-240	-1	0.30	0.92
	ATGGGCC	-240	-41	-240	-21	0.28	1.02
	TAGGC	-180	-21	-120	-21	0.40	0.99
	AATGGG	-220	-1	-160	-1	0.36	0.82
	ATTGGG	-240	-1	-180	-1	0.29	1.08
	ATGGGC	-240	-41	-260	-1	0.75	2.60
	AGGCCC	-240	-41	-240	-21	0.59	2.02
	TTGGGC	-240	-21	-260	-1	0.74	2.37
	AAGCCC	-220	-21	-240	-1	0.59	1.91
	TTCGG	-140	-41	-260	-1	0.60	1.17
	GGCCG	-180	-41	-260	-1	0.72	2.26
	AAGGCC	-200	-21	-220	-1	0.36	0.97
	GGGCCG	-220	-41	-260	-1	0.17	0.87

Table 9 –Motifs defined as overrepresented in U[1:1] and nuclear genes

* Ratio between the number of randomized pairs where the motif is found at its preferential position in both species and the number of randomized pairs of the sample (48,600).

III.D.3. Extended analysis

Once defined the criteria to search for the overrepresented motifs and confirmed the existence of abnormal frequencies of some PLMs within U[1:1] promoters, we extended our analysis with the latest improved version of the *ab initio* detection method developed by V. Bernard. This version, allowed a higher confidence level and an automatic peak window definition for each motif with one base precision. This fact eliminates the window size imposing and gives a higher precision, especially on narrow peaks. Furthermore, new statistical analyses allow a more accurate definition of statistically significant peaks. Additionally, we extended our analyses to include all different subsets of unique genes and use a clean sample of nuclear genes not including the unique genes. In this way, the 14,689 genes with a defined TSS in *Arabidopsis thaliana* were divided in 649 U[1:1], 418 U[1:m], 361 U[1:0]E and 13,261 other nuclear genes, while in *Oryza sativa* the 17,720 genes with defined TSS were splited in 655 U[1:1], 453 U[1:m], 974 U[1:0]E and 15,638 other nuclear genes.

As in previous analysis, input files for the *ab initio* detection program included the 14,689 and 17,720 promoter sequences for which a TSS could be defined in *Arabidopsis thaliana* and *Oryza sativa* respectively. The criteria used to define TSS is, as before, a 5'UTR size of at least 50 bases supported by at least one associated EST or cDNA. The 1000 nucleotides upstream the defined TSS were then retrieved and used as promoter sequences input. Once done, promoters were scanned looking for motifs of sizes ranging from 5 to 8 nucleotides and a SMS higher than 3.

As a result of the new improvements in the *ab initio* detection method, 2,424 PLMs in *Arabidopsis thaliana* and 5,219 PLMs in *Oryza sativa* were defined. The enhanced precision of the detection protocol allows therefore a significant gain on the number of defined motifs. This increase of the number of detected motifs in each species allowed the detection of 1,269 common motifs *i.e.* present in both species, almost tripling the number of common motifs previously observed.

To increase the significance of the obtained results we decided to further filter this 1,269 shared motifs to select only those with similar peak positions in both species. This motif peak conservation would add an additional biological relevance to the obtained motifs as it will imply a similar distance placement for transcription complex formation in both *Arabidopsis thaliana* and *Oryza sativa*. To do so, motif peaks were defined as conserved if their positions in one species include the position in the second species or differ by no more than 20 nucleotides (Figure 28). Such criteria allow a relative flexibility in the definition of similar peak positions that can compensate the increased precision of motif position definition. The number of common motifs with conserved peaks is 937.

Note: The number of common motifs (937) was exactly the same than the number of U[1:1] genes defined in Chapter II. However, this is just a coincidence and it should be remembered that, for the promoter analysis of this chapter, we have only used the 486 U[1:1] genes with a defined TSS.

Extended version of the *ab initio* method for motif detection was capable therefore to define a set of 1,269 common motifs between *Arabidopsis thaliana* and *Oryza sativa* promoters, with 937 of them (74%) sharing similar preferential positions. In order to get the maximum confidence in the results, we selected only these 937 motifs for PLM frequency analysis because we consider that their sequence and position conservations are signs of their functional relevance as TFBS.

III.D.3.a. Overrepresented PLMs and particularities of each sample

Preliminary analysis has shown the existence of a set of overrepresented motifs within U[1:1] promoters despite their sequence independency. In order to obtain a wider view of the possible shared motifs we have extended the analysis by calculating the frequencies of the 937 common motifs between *Arabidopsis thaliana* and *Oryza sativa* with similar peak positions. Moreover, we have detailed the results by independently analysing all the types of unique genes previously defined and eliminating them from the nuclear genes sample used as control. Extended analysis of overrepresented PLMs includes therefore not only a higher number of common motifs but also 4 independent samples of 48,600 randomly paired promoters of each species for U[1:1], U[1:m], U[1:0]E and other nuclear genes. According to the previoulsy defined criteria, motifs were considered as overrepresented if their frequency is higher than the 0.82% cut-off value.

By this way, we have detected 55 overrepresented motifs within U[1:1] promoters. The analysis the other groups of unique genes also showed the presence of overrepresented motifs within their promoters: 52 overrepresented motifs in U[1:m] promoters and 36 overrepresented motifs in U[1:0]E promoters. For the other nuclear genes, 50 overrepresented motifs were detected.

When we compared all the overrepresented motifs found in each group (Figure 29), we found a list of 25 common motifs overrepresented in all the gene groups. Interestingly, such list of motifs presented motifs of 5 nucleotides being in some cases the ‘core’ of motifs found in other groups. Additionally, results confirmed some of our previous observations when studying the unique genes characteristics. On one hand, improved definition of peak positions has allowed the detection of TATA-box related motifs overrepresented in the sample of other nuclear genes. This overrepresentation and the total absence of any related sequences within U[1:1] and U[1:m] promoters confirmed their low frequency of TATA-box motifs previously observed in Chapter II.

Therefore, TATAAA and all other derived TATA-box overrepresented motifs found were only present in other nuclear genes and U[1:0]E gene groups. In the case of U[1:0]E promoters, it is interesting to observe that one of their two specific overrepresented motifs corresponds to the TATATA motif, *i.e.* one of the TATA-box consensus motifs, not found in the group of other nuclear genes. Such dissimilarity could be a sign of a preferential use of this particular version of TATA-box motifs by U[1:0]E genes. On the other hand, the fact that U[1:1] and U[1:m] are particular groups of genes within the nuclear genes was reinforced by the 12 motifs found overrepresented only in these gene groups (Figure 29). These two groups of genes also present 13 overrepresented motifs shared with the other nuclear genes but not with U[1:0]E promoters. This difference suggests that U[1:0]E gene regulation might be assumed by transcription factors dictinct of those of the rest of nuclear genes. In U[1:1] promoters, we observed 4 overrepresented motifs not found overrepresented elsewhere. As two of these motifs are 'derived' from 'core' motifs found in other nuclear genes, only two motifs (CCCATT and ACGGC) seems highly specific to U[1:1] promoters.

Extended analysis confirmed therefore the existence of (i) a set of overrepresented motifs shared within all the nuclear genes, (ii) a low frequency of TATA-box motifs within U[1:1] and U[1:m] promoters, (iii) a low number of putative regulatory motifs within U[1:0]E and, (iv) 4 overrepresented motifs found only within U[1:1] promoters including two specific 'core' motifs (in blank in the Figure 29) .

III.D.3.b. Overrepresented shared motifs in pan-orthologous pairs of U[1:1] promoters.

The different analyses used to detect overrepresented motifs in the different groups of unique genes as well as in the rest of nuclear genes have been performed with the aid of a sample of randomly paired promoters. While this method is necessary to detect overrepresented motifs in these genes for which no direct comparison could be done, U[1:1] genes constitute a group of actual paired promoters, *i.e.* pairs of promoters coming from the pairs of pan-orthologs of *Arabidopsis thaliana* and *Oryza sativa*.

Frequencies of motifs with defined peaks in all nuclear and orthologous promoters were calculated for U[1:1] genes and compared with frequencies in random U[1:1] sample. Results overlap with those previously obtained with random samples since only 3 of the 55 overrepresented motifs found in random U[1:1] promoters are not found in actual pairs. On the other side, we found

27 overrepresented motifs in actual U[1:1] promoter pairs that were not previously detected in random sample.

The 25 overrepresented motifs only found in the promoters of U[1:1] pan-orthologs consist in 68% of the cases on motifs with no similarity with other 'core' motifs and are therefore specific to U[1:1] genes (Figure 30). These motifs may suggest the existence of some kind of shared regulatory mechanisms within U[1:1] genes and increase the probability of a related function despite their sequence independence. Nevertheless, further studies on the number of motifs found in each pair of U[1:1] promoters should be performed in order to study the possible existence of different subgroups of U[1:1] genes with different regulatory speficities.

III.D.3.c. Can we increase the number of overrepresented motifs only detected in pan-orthologous pairs of U[1:1] promoters?

We have confirmed the presence of several overrepresented motifs within the U[1:1] genes promoters. Nevertheless, all the frequency analyses were done with the sequences and preferential motif positions defined by analysing all together the promoters of nuclear genes. It is therefore possible that U[1:1] promoters contain some other overrepresented motifs for which sequence and preferential positions could not be defined in a general analysis of nuclear genes promoters.

To solve this problem, we have analysed the U[1:1] promoter sequences of the 649 *Arabidopsis thaliana* and the 655 *Oryza sativa* genes (having a defined TSS) with the improved *ab initio* detection method. However, the use of such reduced number of promoters to define PLMs decreases the number of PLMs with good confidence level. Indeed, only 197 PLMs in *Arabidopsis thaliana* and 393 PLMs in *Oryza sativa* could be defined. In this case, the number of promoters used to define the motifs was similar in both species while the number of defined motifs in *Oryza sativa* doubled those of *Arabidopsis thaliana*. Due to the size similarity of the samples, their influence on the motif definition could be considered as not relevant but this observation could be explained by the different nucleotide composition of promoters of both species.

The intersection between the 197 and 393 PLMs respectively in *A. thaliana* and *O. sativa* allowed the definition of 81 common motifs. After removing motifs for which the peak position is not conserved in the two species (Figure 28), a final list of 68 common PLMs was obtained. Then, we have calculated the frequencies of each of them within the promoters of actual U[1:1] genes as well as on random samples of U[1:m], U[1:0]E and other nuclear gene promoters. Overrepresented motifs were defined again with the 0.82% cut-off value defined in preliminary analysis. Each list of overrepresented motifs obtained in each group were compared to highlight their differences (Figure

31). There is a high correlation with the previous observed results despite it was possible to observe some apparent 'inconsistencies' such as the not detection of ATGGG motif overrepresentation in the 'other nuclear genes' sample. However such differences were due to the accuracy on motif peak positions which, in the case of ATGGG motif, was higher in motif peaks defined using U[1:1] promoters. While analysis of all nuclear genes defines ATGGG motif peak between positions -175 and -46 in *Arabidopsis thaliana* and positions -155 and -58 in *Oryza sativa*, analysis of U[1:1] promoters defines peak positions with higher precision, placing them between positions -163 and -59 in *Arabidopsis thaliana* and positions -113 and -61 in *Oryza sativa*. Nevertheless, the reduced number of analysed promoters is not expected to increase the accuracy of preferential motif positions.

Despite other small differences found in the intersection with other groups of genes, we have focused our comparison in the 14 motifs found as overrepresented only in the promoters of U[1:1] genes (Figure 31). Such comparison showed that out of the 25 overrepresented motifs previously found in U[1:1] genes (Figure 30), only 3 of them are found overrepresented again when using the list of motifs with preferential positions defined from U[1:1] promoters. Nevertheless, the 11 new overrepresented motifs show, in most of the cases, a high similarity with previously defined motifs as AAAGCCCA and AGCCCAA motifs which can be considered as 'extensions' of the AAGCCCA motif previously found.

Overall, the use of *ab initio* method with U[1:1] promoters in order to looking for new overrepresented motifs is not very efficient and only results in the detection of 'extended' versions of previously defined motifs. However, such 'extended' motifs increase not only the length of the motifs overrepresented in the U[1:1] genes promoters but would also result in a higher specificity of the regulatory motifs used.

III.D.3.d. Preferential positions and lengths of TFBS

As previously described, the preferential positions of the motifs found when analysing U[1:1] promoters could differ from the preferential positions found in the analysis of all nuclear gene promoters. Nevertheless, despite possible differences caused by the different sizes of the input samples used and their consequences on the significance calculation, functional motifs are expected to be conserved around similar positions in both cases (Yamamoto *et al.*, 2007).

Based on this hypothesis, we have plotted in a density graph the 2,424 PLMs in *Arabidopsis thaliana* and the 5,219 PLMs in *Oryza sativa* found from *ab initio* method analysis of all the nuclear genes to display their distribution (Figure 32 A and B). Such figure provides an overview of the

motifs showing (i) a high concentration of motifs between position -50 and TSS in both, *Arabidopsis thaliana* and *Oryza sativa*, (ii) a concentration of motifs in both species between position -100 and –50 with larger peaks ranging from 50 to 200 bases in most of the cases. Overall, this second promoter region motifs form a 'cloud' of points more dispersed than motifs found near the TSS. The repartitions of the 197 PLMs in *Arabidopsis thaliana* and the 393 PLMs in *Oryza sativa* found from *ab initio* method analysis of U[1:1] genes are also similar (Figure 32 C and D). However, in this case the PLMs are in low density concentration around the TSS. Indeed, while in *Oryza sativa* it was possible to observe a slight concentration of motifs near TSS, there were only two spots in *Arabidopsis thaliana*. On the other hand, the motif concentration around positions -100 and -50 could still be observed but again with a lower density and wider peaks in both species as compared to what we observed for all the motifs and genes.

Despite the reduced number of PLMs detected from *ab initio* method analysis of U[1:1] promoters, the comparison with the PLMs detected from *ab initio* analysis of the promoters of all nuclear genes shows a difference in the preferential positions. In U[1:1] promoters, most of the PLMs are localized far from the TSS suggesting a more relaxed position constraints in the transcriptional regulation of the unique genes.

III.D.4. Conclusions

We have successfully used the *ab initio* method developed by V. Bernard to define preferentially located motifs from the analysis of the promoters of all nuclear genes as well as from the promoters of U[1:1] genes. Once done, we have used the sequence and preferential motif positions to extend the search of possible PLMs with abnormal frequencies in promoters of unique gene promoters as previously done for TATA- and TELO-box.

The presence of motifs (with similar positions) was tested on the different groups of unique genes as well as the rest of nuclear genes. This test shows a higher number of overrepresented PLMs within the promoters of U[1:1] genes than in the promoters of other nuclear genes, which was the first evidence of existence of some overrepresented TFBS unique to U[1:1] genes. In fact, a detailed analysis revealed 25 PLMs specific to U[1:1] genes. This result confirms the existence of common regulatory motifs despite the absence of sequence similarity between U[1:1] gene pairs. Nevertheless, the impact of such regulatory similarity in a possible functional relationship of U[1:1] genes remains unknown.

Moreover, observed frequencies of PLMs within the U[1:m] gene pairs confirmed the features similarity previously observed between U[1:1] and U[1:m] genes since almost all the PLMs

overrepresented in U[1:m] promoters were also overrepresented in U[1:1] promoters. Alltoghether, these observations suggest the possibility that most of the U[1:m] genes are due to recent duplications of U[1:1] genes in one of the two species. Similarly, the overrepresented motifs found within the promoters of U[1:0]E genes confirmed that they constitute a particular group of unique genes more similar to the rest of nuclear genes regarding their PLMs frequencies. The almost complete absence of specific regulatory motifs could indicate the independence of each U[1:0]E gene pair in term of regulation.

The presence of ‘core’ and some ‘derived’ versions of motifs suggests the existence of a basic regulatory unit which becomes specific of each gene groups by extending their nucleotidic sequence. Moreover, the fact that motifs defined from promoters of U[1:1] genes present a preferential position between positions -100 and -50 from TSS but not closer, show that these ‘extended’ motifs exhibit a lower positional constraint than the motifs defined from all nuclear gene promoters.

CHAPTER IV

CONCLUSIONS AND PERSPECTIVES

IV. Conclusions and perspectives

IV.A. General conclusion

The work presented in this thesis shows the existence of a set of unique (single-copy) genes within the genome of different plant species with particular characteristics, evolution and promoters. The set of unique genes defined corresponds to a group of genes rarely considered in other studies. In fact, unique genes in each species could be further classified according to the number of orthologous genes found in the other species in U[1:0] (covered or not by ESTs/cDNA), U[1:1] and U[1:m] genes. In this way, while U[1:0] genes have been analysed in some cases (Gutierrez *et al.*, 2004; Vandepoele and Van de Peer, 2005), the study of unique genes conserved in other species constitutes a novelty.

The definition of a set of genes with unambiguous orthology relationship between the different species was the first evident advantage of the study of unique genes. The absence of paralog that could perform the same function and the intron position conservation, which discards convergence, permits a confident knowledge transfer of the biological function between the two species. This unambiguous orthology was the first objective of this thesis since U[1:1] genes were expected to be a perfect set of genes for the promoter analysis in search of conserved motifs. However, the analysis ot the promoters of the othologous genes were delayed by the different characteristics observed within the different subsets of unique genes. As explained in Chapter II, each subset of unique genes was defined based on sequence dissimilarity within and between each species genome. Based on this sequence independence, my preliminary expectations for unique genes supposed that each unique gene would have their own features. Nevertheless, the study of each subset of unique genes showed, on one hand, that the two subsets of unique genes conserved in the two species (U[1:1] and U[1:m] genes) present overall features previously observed in genes with slow evolution rate such as small protein size and high intron density (Carmel *et al.*, 2007). On the other hand, U[1:0]E genes present a particularly short size and a low intron density when compared to the rest of genes. All these particularities highligh that previous observations on U[1:0] genes can not be inferred to the rest of unique genes, but also that U[1:1] and U[1:m] genes present different features than the rest of nuclear genes. Such difference, despite the sequence independence of each unique genes, imply the existence of a selective pressure that conserve not only their features between the two species but which could be also implied in the non-reciprocal local losses between two paralogous duplicated genomic regions. The existence of such selective pressure was supported by the analysis of the 192 U[1:1:1:1] genes defined after the addition of *O. lucimarinus*

and *P. patens* genomes in our study. The presence of 192 U[1:1:1:1] genes among these four species which last ancestor diverged more than thousand million years ago (Zimmer *et al.*, 2007) with similar characteristics as the ones observed in the *Arabidopsis thaliana-Oryza sativa* U[1:1] genes, can be considered as a conclusive proof of such selective pressure. But what is the reason of this selective pressure?

It has been reported that a correlation exists between gene functions and the probability of fixation of duplicated genes after a duplicated event (Krylov *et al.*, 2003; Papp *et al.*, 2003; Blanc and Wolfe, 2004; Maere *et al.*, 2005). However, such correlation was difficult to analyze in my case because only 20-30% of unique genes were covered by a GO annotation, remaining the function of the rest unknown. Nevertheless, despite this lack of coverage, in the case of U[1:0] genes the analysis of the 105 AtU[1:0]E with annotated function detected the presence of different known peptide phytohormones (Farrokhi *et al.*, 2008) such as CLAVATA3 and 5 related peptides, POLARIS, 3 PROPEP, RALF and N Hydroxyprolin-rich glycoprotein coding genes. This finding joined to the observed features of U[1:0] genes such as very short sequence, low intron density and a relatively high percentage of genes coding for proteins targeted to the endoplasmatic reticulum suggests that U[1:0] genes are implied in the coding of phytohormones precursors of secreted peptides with probably regulatory role in defence or non-defence functions (Farrokhi *et al.*, 2008). The predicted function of U[1:0] would imply therefore an important aspect. Due to their definition as species-specific, the origin of U[1:0] genes is expected to have occurred probably after radiation event and be implied in the particular phenotypes of each species. Nevertheless, many of the described peptide phytohormones within AtU[1:0] such as CLAVATA3 are also present in *Oryza sativa*. In fact some peptide phytohormones may be clustered based on short motif conservation characterised by only 12 residues while the remaining parts of the propeptides are highly divergent. This characteristic mades their detection by sequence similarity difficult and implies that they may be present in other plant species and that they may have been duplicated during evolution but we are not capable to detect it.

The analysis of U[1:1] and U[1:1:1:1] genes covered by functional annotations could not highlight any particular tendency in their functions. Nevertheless, their features as well as the observed abnormal occurrence of TATA-and TELO-box in U[1:1] and U[1:1:1:1] promoters suggests that they could be linked to critical housekeeping functions such as protein catabolism and synthesis, RNA processing or DNA repair. This critical housekeeping function would explain not only their conservation in different photosynthetic species but also their reduced fixation probabilities during duplication and their maintenance as single copy genes in the different species due to the necessity of strict regulate them, maintain their stochiometry and possibly avoid an

excessive use of limited resources. Regardless of the U[1:1] and U[1:1:1:1] gene functions and their impact on keeping their uniqueness, what I can conclude from U[1:1] genes conservation of protein sizes, transcription levels and sequence conservation (dN/dS) is that gene loss occurred rapidly after the duplication. Such quick gene loss would therefore reinforce the idea of a probably important conserved function. This reduced time of divergence discards possible paralog noise in our characterization of pan-orthologs as at the time of gene loss they would not be differentiable from actual orthologs.

Overall, despite the definition of U[1:1] and U[1:1:1:1] genes was based only on sequence similarity between each pair or quadruplet of genes and was, therefore, independent between them, their features and the not fixation of duplicated genes after a duplicated event suggest that unique gene functions could be somewhat similar. This possibility increased the interest of 'phylogenetic footprinting' analysis of U[1:1] genes, *i.e.* unambiguous orthologs, as well as the analyse of promoters of other unique gene subgroups looking for other possible particularities. These two points were analysed separately with different success rates. First, a 'phylogenetic footprinting' study was performed by analysing different U[1:1] pairs of genes exhaustively (DECOMO) and with a Gibbs sampling method (MotifSampler) to detect conserved motifs between U[1:1] gene pairs. The obtained results with these two methods were then filtered using a position constraint filter and a 'maximum number of ordered motifs' filter. Unfortunately, these two methods and their filters, which proved their capacity to highlight differences between orthologous and non-orthologous promoters of an *Arabidopsis thaliana-Raphanus sativus* control sample, were unable to find differences between the promoters of U[1:1] genes and randomly paired genes of *Arabidopsis thaliana-Oryza sativa* or other closer species. The results show that the divergence time between the different species was high enough to remove all the traces of ancestral sequence and hide conserved functional motifs within a variable number of 'background noise' motifs. Second, I used an *ab initio* method developed by Virginie Bernard in the framework of her thesis to define relevant motifs within the promoters of all nuclear genes as well as within the promoters of U[1:1] genes. Unlike first 'phylogenetic footprinting' analysis, this method allowed me to work with a set of motifs with significant position preferences and filter out many 'not significant' motifs. Moreover, the analysis of motif frequencies without the need of defined unique genes pairs allowed me to analyse all the subsets of unique genes in a similar way as previously done to detect abnormal TATA- and TELO-box frequencies. This approximation permits to prove the existence of overrepresented motifs in each subset of unique genes. Moreover, the comparison of motif lists found in the different subsets of unique genes confirmed the previous differences observed. On one hand, U[1:0]E genes presented within their promoters an overrepresentation of many of the motifs

found in the not unique genes, including TATA-like motifs, but not many of the motifs present in other nuclear, U[1:1] and U[1:m] genes. The lack of overrepresentation for many other motifs suggests that U[1:0]E genes may contain a variable number of orphan genes appeared after radiation that could present a particular regulation system. On the other hand, the similarities found between U[1:1] and U[1:m] sustain the similar features previously observed between these two types of unique genes, *e.g.* short size and high intron frequency. Nevertheless, the most important results obtained was the definition of a set of motifs only overrepresented within the promoters of U[1:1] genes. This list of overrepresented motifs found within the promoters of actual U[1:1] genes in two different species suggests that despite the definition of U[1:1] was based only on sequence similarity between each pair of genes these genes could perform related functions and be regulated by the same regulatory motifs. The fact of finding similar regulatory motifs sustains the probable housekeeping function of U[1:1] genes as most of the transcription profiles of 'housekeeping' genes have been described as constitutively transcribed at different levels among cell types and shut down in response to extreme environmental conditions such as heat shock or starvation (Pirkkala *et al.*, 2001)

IV.B. Extension of 'phylogenetic footprinting' analysis

Despite the detection of particular motif overrepresentation within the promoters of U[1:1] genes, I can not conclude that they are all regulated in a similar way. This lack of conclusion is due to the fact that I quantified the number of overrepresented motifs overall but not in a pair by pair basis. Consequently, it could be argued that the overrepresented motifs found are only the contribution of a fraction of U[1:1] genes presenting most of the detected promoters while the rest of U[1:1] genes would present motifs similar to those found in not unique genes. This possibility is supported by the fact that unique genes could have been originated at different points along evolution and followed different conservation schemas (Figure 33). The phylogenetic profiles of conserved single-copy genes and the predicted subcellular locations of the corresponding proteins, have proved these different origins by further dividing the set of U[1:1] genes in different subgroups. In this way, while U[1:1] genes conserved in bacteria, plants and metazoan should be conserved since ancestral prokaryotic cell and would be more likely linked to critical housekeeping functions, whereas U[1:1] genes specific to plants should have appeared after radiation from metazoan and would probably be related with particular functions of plants such as photosynthesis or cell wall. Each of these subgroups of U[1:1] genes as well as the fraction of U[1:1] genes with evidences of an ancestral horizontal gene transfer or those U[1:1] genes only present in plants and

metazoan (conserved from an ancestral eukaryotic cell) can be expected to be regulated by their own particular set of regulatory motifs.

To test this hypothesis and highlight the possible regulation similarities between and within the different types of U[1:1] genes it would be necessary to extended the analysis based on the motifs detected with *ab initio* method differentiating each of the different subgroups of U[1:1] genes. Other more classical 'phylogenetic footprinting' analysis based only on sequence comparison but without considering significant preferential positions of the motifs would probably fail even if I can add more species to the analysis (as observed with the MotifSampler analysis of U[1:1:1:1] genes) because the species radiation occurred too long time ago that they might no longer have shared regulatory motifs (Pan *et al.*, 2008). Furthermore, I could extend the *ab initio* analysis of conserved motifs to include also the regions place within the 5' UTR, downstream the TSS. These regions have not been analysed during this thesis to reduce calculations workload and avoid possible misinterpretations in the number of conserved motifs calculated with DECOMO due to the possible presence of microsatelites within 5'UTR (Lobo-Menendez *et al.*, 2004). Nevertheless, several known transcription factors binding sites placements have been described after the TSS, *e.g.* the DPE motif. The use of *ab initio* method combined with a larger promoter sequence can therefore highlight new motif overrepresentations within the different U[1:1] gene subsets. Another important point to take into account that has not been considered in this thesis would consist on the analysis of complementary strand of the studied genes. Indeed, I have not considered complementary strand to reduce the number of obtained motifs and concentrate on the possible differences between ortholog and non-orhtolog genes. This decision permitted to simplify my analysis while remaining significant enough to discard DECOMO and MotifSampler for the analysis of such distant species.

The use of 5'UTR and complementary strand to extend the promoter analyis using *ab initio* method is expected to give a wider view of the overrepresented motifs in each particular subset of unique and U[1:1] genes. However, this extended analysis of the promoter presents one main flaw for the analysis of overrepresented motifs in each U[1:1] genes subset. Indeed, as observed when analysing the motifs defined by *ab initio* method from all the nuclear gene promoters and U[1:1] gene promoters, they can vary not only in the number and sequence of the motifs defined but also in their preferential positions. This way, the size of the sample used is very important for the definition of motifs with significant preferential position since it can influence the statistics used to determine the significance of the hits. As a consequence of this size importance, the extended analysis for the detection of particular overrepresented motifs in each of the U[1:1] genes subsets would be limited to the analysis of those motifs defined from the analysis of all nuclear genes or U[1:1] promoters.

Regardless of the possible limitations for the definition of a motif set for each subtype of U[1:1] genes, it would be interesting to include two other points in the extended analysis. The first point is the possibility to analyse 'underrepresented' motifs within each group of unique genes. It has been observed that some of the motifs analysed during the promoter analyses were found in a very low rate and it was even possible that they were not present at all within a group of unique genes. Therefore, it would be possible in detailed analysis to define and quantify the absent motifs and use them in the second point. The second point that would be interesting to test with both the overrepresented and absent motifs is their search within the literature and databases in order to establish how many of them are known and, if it is the case in which functions are they related. This approach would be specially helpful to infer functional information of U[1:m] and U[1:1] (including their subsets) genes for which no functional information was obtained from GO annotation. These two last points are probably the most easy to achieve as they would need no further analysis of the promoters but a study of the already obtained results.

IV.C. Exploitation of transcriptome analysis to infer functional information to unique genes

The number of data available from transcriptome analysis has grown exponentially in the last years with projects like CATMA (Complete *Arabidopsis* Transcriptome MicroArray) that has currently produced 5,906 hybridized samples (Gagnot *et al.*, 2008), or Genevestigator which, after the recent addition of rice (September 10th, 2008), contains the data of 20,900 microarrays from 6 different species (Zimmermann *et al.*, 2004).

Unfortunately, during the analysis of U[1:1] genes all this information could not be used to analyse the level of expression of the genes on each species and study their possible correlation. Indeed, the problem that I encountered was not the lack of transcriptome analysis *per se* but the impossibility to compare them in the same conditions in both species. While this impossibility of comparing transcription in both species has forced me to apply another approximation to study the relative level of transcription by using the number of associated ESTs to the gene in each species (Chapter II), it is possible to exploit current transcriptome analysis to infer functional information to unique gene (Horan *et al.*, 2008). In this way, I expect that a comparison of expression profiles between the unique genes and the rest of nuclear genes in different tissues could establish some kind of similarities that can link unique genes with unknown function to a group of genes implied in

a determined biological or biochemical response. Such analysis would need the use of trancriptome data backed up with strong statistical analysis to confirm the obtained results.

IV.D. Description of novel signalling and disease related genes in U[1:0] genes

One of the most interesting groups of unique genes to analyse is the group of orphan genes or U[1:0]. Orphan genes between close related *Oryza* species have shown their role in defence roles (Sakai and Itoh, 2008), a role reinforced in U[1:0] genes due to the small size of their coding protein. Such small proteins have been reported to code for peptide phytohormones involved in signalling roles in defence or non-defence functions (Farrokhi *et al.*, 2008). We can expect therefore that many of the U[1:0] genes defined with my protocol are related to similar functions. However, the transcriptome analysis proposed here to assign them a functional role may not be the most appropriate.

One of the main problems that I expect to encounter during transcriptome analysis of U[1:0] genes is the fact that transcriptome analysis show the actual state of genes activate within a tissue or organ under a certain conditions. As this condition is usually based on 'normal' condition for the plants, if U[1:0] genes are related with immune response as suggested by their characteristics, I expect to find them widely inactive. In such situation I would need to study the transcriptome of plants under stress conditions or infected by diseases of diverse origin to have a wide view of the possible conditions on which potentially defence related U[1:0] genes could be involved. In this sense, the amount of experience data currently available covers most of the stress and infection conditions and is rapidly increasing which opens the door to the analysis of U[1:0] genes transcriptome. Nevertheless, another option for the potential description of novel signalling and disease related functions of U[1:0] genes should be considered. Indeed all the methods and conclusions described in this thesis are based on informatics analysis of biological data that are the only way to analyse large amounts of data obtained by high-throughput techniques but which results are not as conclusive as a 'wet-lab' approach. On the contrary, 'wet-lab' analyses are basically limited by resources and time related issues. However, in the case of U[1:0] genes, their analyses have allowed me to defined a small group of 13 pro- and gly- rich proteins with high probabilities of implication in novel signalling or disease related functions. Therefore, an actual 'wet-lab' analysis of such genes could be feasible and produce very interesting results. Similar 'wet-lab' analysis could be considered for the 35 U[1:1:1:1] genes found in all the phyla under study. Despite

that in this case the amount of work would be greatly increased, the expected results could provide key information about those genes with an ancestral function not yet described.

Finally, it should be considered that many small predicted genes have been previously discarded during annotation of species genome because they showed no similarities with any other known sequence (Termier and Kalogeropoulos, 1996). However, recent studies have showed that many of these small predicted genes are actually functional (Harrison *et al.*, 2003; Galindo *et al.*, 2007) which will imply the necessity to revise many of the previously discarded genes. This situation will increase the number of U[1:0] genes available for study and may provide new strong candidates for novel signalling or disease related roles.

IV.E. Possible recent duplication of U[1:m] genes

Along the work included in this thesis, one of the groups of unique genes for which I could provide less information are the U[1:m] genes. While U[1:0] genes present characteristics of genes related to signalling and defences roles, and U[1:1] genes characteristics and conservation along phylogeny in single-copy suggest a critical housekeeping function, the U[1:m] genes are just described as having similar characteristics as U[1:1] genes. But what are their functions and origins? Indeed, I expect that, as U[1:1] genes, they could be related to housekeeping functions but probably different that those done by U[1:1] genes due to the differences observed in the over-represented motifs within their respective promoters. Nevertheless, as explained for U[1:1] genes, U[1:m] genes are probably a mix of genes with different origins. One of the possibilities in which I am particularly interested is their possible recent duplication in one of the species following a local or segmental duplication event which has not yet been lost. If this is the case, I expect that the 'm' value in U[1:m] genes would be frequently a '2' and that both genes would be almost identical. Under these premises, it would be interesting to analyse U[1:m] genes to calculate the incidence of recent duplications on them and, even, the mean dN and dS values of genes in such situation. Overall, the work presented in this thesis shows that unique (single-copy) genes have characteristics different to the rest of nuclear genes but similar within each type of unique gene despite their sequence independence. The analyses of these characteristics, as well as their conservation as unique gene along evolution despite WGD, have given some clues about their functions despite their lack of GO annotations. Nevertheless, I expect that the future analysis of the overrepresented motifs that I have found in their promoters and the possible underrepresented motifs, as well as the analysis of the transcriptome can provide further information about the biological function of these genes for which I presume an ancient (but not yet annotated) critical function.

CHAPTER V

ANNEXES

V. Annexes

V.A. Figure index

Page

V.B. Table index

V.C. Abbreviations

2R	2 Rounds (of duplication)
aa	aminoacid
AGI	Arabidopsis Genome Initiative
bp	base pair
cDNA	complementary DNA
CDS	Coding Sequence
DNA	DesoxyriboNucleic Acid
DDC	Duplication-Degeneration-Complementation
DMI	Dobzhansky-Muller Incompatibilities
dN	Non-synonymous substitution
DPE	Downstream Promoter Element
dS	Synonymous substitution
EM	Expectation Maximization
ER	Endoplamic Reticulum
EST	Expressed Sequence Tag
GFF	General Feature Format
GO	Gene Ontology
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
Inr	Initiation Element
Mb	Megabase
MYA	Million Years Ago
Myr	Million Years
Nb	Number
ORF	Open Reading Frame
PLM	Preferentially Located Motif
POF	Protein with Obscure Features
PPI	Protein-Protein Interaction
PWM	Position Weight Matrix
RBH	Reciprocal Best Hit
SMS	Score Maximal Square
TAF	TBP-Associated Factors
TBP	TATA-Binding Protein
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Starting Site
U[1:0]	Unique gene in one species with no homologs in the other species
U[1:0]E	U[1:0] gene with Expression evidence
U[1:0]NE	U[1:0] gene with No Expression evidence
U[1:1]	Unique gene in one species with only one homolog in the other species
U[1:m]	Unique gene in one species with many homologs in the other species
WGD	Whole Genome Duplication

V.D. Resumen en español

Objetivos

Los objetivos de mi tesis han sido:

(i) poner de manifiesto la presión de selección que se ejercen sobres los genes ortólogos en plantas

(ii) describir las características estructurales y funcionales que comparten, especialmente a nivel del promotor, y que puedan usarse en una aproximación por huella filogenética asi como determinar los limites de la misma.

Para apoyar el estudio en unas relaciones de ortología lo menos ambiguas posibles, el estudio se ha realizado en genes 'únicos' en plantas.

Introducción

En la literatura existen dos definiciones de genes únicos diferentes e independientes de la noción de ortología. La primera definición esta basada en la secuencia. Según esta primera defínción, los genes únicos en un determinado genoma, no poseen ninguna similitud detectable con el resto de genes del genoma aunque puedan poseer una función similar a la de otros genes. La segunda definición de genes únicos está basada en la función. Sin embargo, esta definición es difícil de utilizar ya que la información funcional validada de manera experimental es a menudo insuficiente, particularmente en lo que concierne a la función biológica.

En esta tesis, se ha utilizado la definición de genes únicos basada en la secuencia. Gracias a la comparación de secuencias he podido definir un grupo de genes únicos en los genomas de diferentes plantas. Una vez hecho esto mi objetivo ha sido establecer las características comunes de estos genes utilizando para ello un método que me permitiera definir los genes únicos primando la eliminación del mayor número de falsos positivos sobre la exhaustividad.

Desde el punto de vista evolutivo, la existencia de genes únicos plantea una cuestión interesante. A lo largo de la evolución las plantas han conocido diversos episodios de duplicación, completa o parcial, de sus genomas y de alopoliploidia que deberían hacer muy difícil la presencia de genes únicos en los genomas actuales. Sin embargo, el retorno de los genes poliploides ha un estado diploide, bien sea por divergencia de los genes duplicados o por una deleción masiva de genes, parece la norma general. Los procesos de duplicación seguidos de divergencia o deleciones permiten, por un lado, la formación de largas familias de parálogos en el interior de un genoma y, por otra parte, la creación de una cantidad variable de genes únicos. En la introducción de la tesis, el

rol de la secuencia evolutiva 'duplicación-retorno al estado único' se ha discutido desde el punto de vista de la aparición de nuevas funciones y de la especiación.

Los genes únicos observados en un genoma pueden, o no, haber sido conservados en forma de homólogos en otros genomas. En el caso de que un gen único esté conservado en alguna otra especie, éste puede encontrarse asimismo en estado único o bien formando una familia de parálogos. Las diferentes categorías de genes únicos conservados en ocasiones en especies filogenéticamente muy alejadas, son transversales a las diferentes clases de homólogos. La hipótesis inicial es por tanto que los genes conservados como únicos en dos especies bastante alejadas filogenéticamente como son *A. thaliana* y *O. sativa* (cuya divergencia se estima en alrededor de 150 millones de años) son muy probablemente genes ortólogos.

La disponibilidad de secuencias completas de diversos genomas de plantas pertenecientes a especies que van desde un alga monocelular a angiospermas, me han permitido realizar un amplio estudio de genómica comparativa para caracterizarar los genes únicos en plantas. En una primera aproximación, se caracterizaron los genes únicos de *A. thaliana* y de *O. sativa* y se definio su conservación en las dos especies. Seguidamente, se extendio la aproximación por genómica comparativa a otros genomas de plantas cuya especiación es anterior a la separación de monocotiledóneas y eudicotiledóneas.

Los genes únicos forman 3 grupos estructuralmente, funcionalmente y evolutivamente diferentes.

Un protocolo estricto para eliminar todos los miembros de familias, basado en BLAST y en los motivos conservados definidos por PFAM, fue aplicado sobre los proteomas de *A. thaliana* y de *O. sativa*. Este protocolo me permitió establecer una lista de genes únicos para cada una de las dos especies. El número de genes únicos es de 2,570 en *A. thaliana* y de 8,041 en *O. sativa*. El estudio de la localización cromosomica de los genes únicos ha mostrado que su unicidad no se explica por una presencia más elevada en las regiones cromosomicas sin restos visibles de duplicación segmental. En efecto, estas regiones que cubren alrededor de un 16% del genoma de *A. thaliana* no contienen una mayor cantidad de genes únicos que aquellas regiones para las que existen pruebas de su duplicación. La mayoría de genes únicos ha sufrido por tanto duplicaciones, de igual manera que el resto de genes, pero tras la duplicación estos genes han retornado a un estado de unicidad contrariamente al resto de genes cuya duplicación seguida de diversificación ha conducido a la formación de familias multigénicas.

Seguidamente crucé por comparación de secuencias las listas de genes únicos en las 2 especies para identificar los pares de genes únicos conservados en ambas plantas. Los 937 pares de genes definidos de esta manera me han permitido poner de manifiesto las correlaciones a nivel de la talla de la proteína y de los niveles de expresión entre los dos genes de cada par. Los genes únicos conservados en *A. thaliana* y en *O. sativa* están caracterizados por un número anormalmente elevado de intrones y codifican proteínas de una talla más pequeña que el resto de genes nucleares. Estas características son propias de genes de evolución lenta. Un análisis global de cuatro motivos reguladores conocidos ha puesto en evidencia una sub-representación de la caja TATA y una sobre-representación de la caja TELO en los promotores de los genes únicos. El conjunto de genes únicos define por tanto una clase funcional particular. También he mostrado que el nivel de transcripción de los pares de genes únicos, estimado gracias al número de secuencias transcritas conocidas, esta positivamente correlacionado con el nivel de similaridad de sus secuencias. Existe además una conservación significativa en el número y la posición de los intrones en los pares de genes únicos de *A. thaliana* y de *O. sativa*. Cabe remarcar que estos análisis se han realizado exclusivamente sobre aquellos pares de genes únicos cuya estructura intron-exon ha sido validada experimentalmente por la presencia de transcritos. El conjunto de estas observaciones, algunas de ellas independientes de la conservación de la secuencia, sugiere que estos genes únicos conservados en *A. thaliana* y *O. sativa* son ortólogos *strictu sensu*.

Las particularidades de los genes únicos en comparación al conjunto de los genes nucleares me condujeron a interesarme por su conservación en los genomas de otros representantes del reino vegetal. A pesar de la rigurosidad de nuestra selección de genes únicos, pude identificar 192 genes conservados como únicos en los genomas de *A. thaliana*, *O. sativa*, el musgo *P. patens* y el alga unicelular *O. lucimarinus*. A modo comparativo, una conservación aleatoria de genes únicos desde *O. lucimarinus* hasta *A. thaliana* no superaría teóricamente los 8 o 9 genes. Estos genes únicos altamente conservados tienen una función bioquímica desconocida. Además, estos genes han permanecido únicos a pesar de los diversos eventos de duplicación, global, local y segmental, que han tenido lugar desde la especiación de las algas verdes hace más de 700 millones de años. Por tanto, podemos concluir que una fuerte presión de selección ha mantenido como únicos un número importante de genes. El perfil filogenético de las proteínas codificadas por estos genes únicos y el estudio de su localización subcelular predicha, indican que alrededor de un 20% de estos genes podría tener un origen bacteriano o provenir de una transferencia de ADN del cloroplasto al núcleo. El mantenimiento de la estequiometria entre los productos de estos genes y de las proteínas con las que formarían complejos funcionales en los cloroplastos explicaría su mantenimiento como únicos dentro del genoma nuclear tras la transferencia.

Contrariamente a lo esperado, tanto los genes únicos específicos de *A. thaliana* como los genes únicos específicos de *O. sativa* presentan unas características comunes. Estos genes son diferentes de los genes únicos conservados puesto que, de media, son particularmente cortos, pobres en intrones y codifican frecuentemente proteínas dirigidas al retículo endoplasmático. Solamente un 25% de estos genes están asociados a una anotación GO y entre ellos se encuentran las diferentes clases de péptidos a los cuales se atribuye un rol hormonal. Estos péptidos cortos provienen de la maduración post-traduccional de pro-péptidos que, dentro de una familia funcional dada como p.e. los CLE, solo tienen en común un número muy pequeño de aminoácidos que pueden ser reconocidos como homólogos. Estos genes únicos específicos de una especie parecen por tanto constituir una clase muy particular de genes para los que la perdida de una copia tras un evento de duplicación estaría favorecida bien por una deleción, bien por una evolución rápida de la secuencia asociada a la especiación.

Los promotores de los genes únicos conservados contienen un pequeño número de motivos sobre-representados.

Como resultado de este estudio es possible definir tres clases de genes únicos presentes en *A. thaliana* y *O. sativa*: genes únicos específicos de una especie, pares de genes ortólogos únicos en ambas especies y grupos de genes que incluyen un gen único en una de las especies y diversos homólogos en la otra especie. Una vez definidos los grupos de genes, estudié la composición en secuencias potencialmente reguladoras en sus promotores. Para ello utilizé una aproximación basada en la noción de huella filogenética para intentar evidenciar una presencia preferencial de ciertas secuencias en los pares de genes ortólogos únicos en *A. thaliana* y *O. sativa*. Para ello solamente utilicé los genes cuyos puntos de iniciación de la transcripción han sido definidos de manera experimental, lo que corresponde a un poco más de la mitad de los genes en las dos especies. Tras probar una toda una serie de métodos disponibles, MotifSampler se mostró como el método que respondía mejor a las condiciones deseadas: rapidez, posibilidad de ser utilizado de manera local, facilidad para analizar de manera automática el resultado y capacidad de reconocer las secuencias introducidas en un set de prueba.

Igualmente desarrollé un método, DECOMO (por DEscription of COnserved MOtifs), que permite una descripción exhaustiva de los motivos conservados en los dos promotores de cada par de genes únicos conservados. Para disminuir el número de posibles falsos positivos, es decir, para aumentar la probabilidad de que un motivo detectado sea un motivo funcional, asocié al programa un cierto número de filtros como: el tamaño de los motivos, el orden de los motivos sucesivos, la

presencia de motivos en la misma región en los dos promotores y las frecuencias excepcionales. Ni MotifSampler ni DECOMO pudieron evidenciar un aumento en el número de motivos presentes en los genes únicos conservados en *A. thaliana* y *O. sativa* en comparación con el número de motivos detectados en pares aleatorios de genes. Las pruebas con pares de plantas filogenéticamente más próximas, *Arabidoposis* y rábano o álamo y viña, mostraron un mayor número de motivos comunes entre los genes únicos conservados. Sin embargo, este efecto podría ser debido a una mayor conservación global entre las secuencias promotoras y no a la conservación de motivos *stricto sensu*.

A consecuencia de estos resultados, introduje entonces un método de predicción *ab initio* de motivos que podrían ser puntos potenciales de fijación de la transcripción. El método utilizado está basado en la localización preferencial de los motivos en los promotores. Este método ha sido implementado en el seno del laboratorio por V. Bernard durante su tesis (actualmente en curso). Existen 937 motivos con una situados preferencialmente en la misma posición en el promotor de *A. thaliana* y de *O. sativa*.

La comparación de los diferentes grupos de genes con este método confirmó que:

(i) existen motivos sobre-representados en el conjunto de genes nucleares además de:

(ii) detectar menos motivos en los promotores de los genes únicos no conservados,

(iii) que los promotores de los genes únicos conservados presentan una sub-representación de TATA-box,

(iv) que existen 13 motivos sobre-representados comunes en los genes únicos conservados entre *A. thaliana* y *O. sativa*,

(v) y 25 motivos sobre-representados únicamente en los genes únicos.

Conclusiones

Antes de este trabajo, los diversos estudios publicados habían sido dedicados al analisis de los genes únicos específicos de una especie. La originalidad de esta tesis ha sido estudiar en paralelo, por una parte, las características estructurales y funcionales y, por otra parte, la conservación en el curso de la evolución de las diferentes clases de genes únicos. Mis resultados demuestran que los genes únicos conservados y los genes únicos no conservados no comparten las mismas características tanto a nivel de estructura de los genes y de la talla de la proteína codificada por estos genes, como a nivel funcional y a nivel de presencia de motivos en sus promotores. Así, por una parte, los genes únicos específicos de una especie son preferencialmente genes que codifican péptidos excretados e implicados en funciones de regulación que pueden estar ligadas a mecanismos de defensa o de diferenciación celular. Por otra parte, los genes únicos conservados poseen características que están

asociadas a genes implicados en funciones de metabolismo de base de las células y que evolucionan lentamente. Por otro lado, mis observaciones indican que, generalmente, estos genes una vez duplicados vuelven rápidamente a un estado de unicidad. En consecuencia, el carácter único y la relación de ortología estricta son mantenidos durante la evolución lo que sugiere una fuerte desventaja selectiva a la existencia simultánea de dos copias poco divergentes de estos genes.

Perspectivas

Para concluir este trabajo me gustaria proponer ciertas indicaciones para futuros estudios. (i) El análisis de los promotores de las diferentes clases de genes únicos por el método *ab initio* basado en restricciones topológicas y su conservación podría ser extendido a la región 5'UTR conocida por contener secuencias reguladoras de la transcripción. (ii) La utilización de los datos del transcriptoma de *A. thaliana* y de *O. sativa* para comparar la expresión de los genes únicos conservados, así como también para explorar la naturaleza de los genes únicos específicos de una especie y su implicación en las vías de señalización y de respuesta a factores bióticos. Esta aproximación constituiría una primera etapa para la validación experimental de la función de las 13 proteínas ricas en prolina y glicina codificadas por genes únicos específicos de una especie, así como de los 35 genes únicos conservados en plantas pero también presentes en una amplia variedad de organismos. (iii) Las razones de la existencia de dos categorías de genes únicos conservados, aquellos conservados únicos y aquellos que presentan varios homólogos en uno de los genomas, podrían buscarse en la edad de la duplicación o en la fuerza de la presión de selección asociada la función de estos genes.

V.E. Resumé en français

Objectifs

Les objectifs de ma thèse étaient :

(i) de mettre en évidence les pressions de sélection qui s'exercent sur les gènes orthologues chez les végétaux et

(ii) de cerner les caractéristiques structurales et fonctionnelles qu'ils partagent, notamment au niveau de leur promoteur pour mieux définir une démarche d'empreinte phylogénétique et en cerner les limites.

Pour espérer s'appuyer sur des relations d'orthologie les moins ambiguës possibles l'étude a été centrée sur les gènes 'uniques'.

Introduction

Dans la littérature, deux définitions différentes mais toutes deux indépendantes de la notion d'orthologie, ont été appliquées aux gènes uniques. La première définition est basée sur la séquence. Dans ce cas, les gènes uniques, dans un génome donné, n'ont de similarité détectable avec aucun autre gène du même génome même s'ils peuvent avoir une fonction proche de celle d'autres gènes. La deuxième définition des gènes uniques est basée sur la fonction. Cette notion est cependant difficile à manier car l'information fonctionnelle, expérimentalement validée, est souvent insuffisante, particulièrement en ce qui concerne la fonction biologique.

Dans cette thèse, c'est la définition des gènes uniques basée sur la séquence qui a été retenue. Opérationnellement, c'est donc la comparaison de séquences qui a permis de désigner un groupe de gènes uniques dans les génomes de différentes plantes. Mon but étant d'établir les caractéristiques partagées par les gènes uniques, l'approche utilisée pour sélectionner les gènes uniques a pour objectif d'exclure autant que possible les faux positifs plutôt que de viser l'exhaustivité.

Du point de vue évolutif, l'existence de gènes uniques pose une question intéressante. En effet, les végétaux connaissent des évènements de duplication complète ou partielle de leurs génomes et d'allopolyploïdie qui devraient rendre rare la présence de gènes uniques dans les génomes actuels. Cependant, le retour des génomes polyploïdes vers l'état diploïde, soit par divergence des gènes dupliqués soit par délétion massive de gènes, semble être la règle. Les processus de duplications suivies de divergence ou de délétions permettent, d'une part, la formation de larges familles de paralogues à l'intérieur d'un génome et, d'autre part, la création des gènes

uniques visibles en nombre variable dans les génomes actuels. Dans l'introduction de la thèse, le rôle de la séquence évolutive 'duplication-retour à l'état unique' est discuté dans le cadre de l'apparition de fonctions nouvelles et de la spéciation.

Les gènes uniques observés dans un génome peuvent ou non avoir été conservés sous forme d'homologues dans d'autres génomes. Dans le cas où il y a conservation dans une autre espèce d'un gène unique, cela peut se faire soit sous la forme d'un gène homologue également unique soit sous la forme d'une famille de paralogues. Les différentes catégories de gènes uniques conservés, parfois dans des espèces très éloignées phylogénétiquement, sont transversales aux différentes classes d'homologues. L'hypothèse de départ est cependant que les gènes uniques mais conservés uniques entre deux espèces assez éloignées phylogénétiquement comme *A. thaliana* et *O. sativa* (dont la divergence date d'environ 150 millions d'années) sont très probablement des gènes orthologues.

La disponibilité des séquences complètes de plusieurs génomes de plantes appartenant à des espèces allant d'une algue monocellulaire aux angiospermes, m'a permis de réaliser une large étude de génomique comparative pour caractériser les gènes uniques chez les plantes. Dans une première approche, les gènes uniques d'*A. thaliana* et d'*O. sativa* ont été caractérisés et leur conservation entre ces deux espèces a été définie. Ensuite, l'approche de génomique comparative a été élargie à d'autres génomes de plantes dont la spéciation était antérieure à la séparation des monocotylédones et des eudicotylédones.

Les gènes uniques forment 3 groupes structurellement, fonctionnellement et évolutivement distincts

Un protocole stringent, s'appuyant sur BLAST et les motifs conservés définis par la ressource PFAM pour éliminer tous les membres de familles, a été appliqué sur les protéomes d'*A. thaliana* et d'*O. sativa*. Les listes de gènes uniques ont été établies pour ces 2 espèces. Le nombre de gènes uniques est de 2,570 chez *A. thaliana* et de 8,041 chez *O. sativa*. L'étude de la localisation chromosomique des gènes uniques a montré que leur unicité ne s'explique pas par leur sur-représentation dans des régions chromosomiques sans trace encore visible de duplication segmentale. En effet, ces régions qui couvrent environ 16% du génome d'*A. thaliana* ne contiennent pas plus de gènes uniques que les régions chromosomiques pour lesquelles il existe des évidences de duplication. La majorité des gènes uniques a donc bien subi des duplications, au même titre que l'ensemble des autres gènes, mais après duplication ces gènes sont revenus à l'état unique contrairement à d'autres gènes dont la duplication suivie de diversification a conduit à la formation de familles multigéniques.

J'ai ensuite croisé, par comparaison de séquences, les listes de gènes uniques des 2 espèces et identifié les paires de gènes uniques conservés entre ces deux plantes. Les 937 paires ainsi définies ont permis de mettre en évidence des corrélations au niveau de la taille de la protéine et des niveaux d'expression entre les deux gènes d'une même paire. Les gènes uniques conservés chez *A. thaliana* et *O. sativa* sont caractérisés par un nombre anormalement élevé d'introns et ils codent pour des protéines de plus petite taille que l'ensemble des gènes nucléaires. Ces caractéristiques sont celles de gènes dont l'évolution est lente. Une analyse globale de motifs régulateurs connus a mis en évidence une sous-représentation de la TATA-box et une sur-représentation de la TELO-box dans les promoteurs des gènes uniques. L'ensemble des gènes uniques défini donc une classe fonctionnelle particulière. J'ai également montré que le niveau de transcription des couples de gènes uniques, estimé à l'aide du nombre de séquences transcrites connues, est positivement corrélé à leur similarité de séquences. Il existe par ailleurs une conservation significative du nombre et de la position des introns entre les gènes d'*A. thaliana* et d'*O. sativa*. Notons que ces analysis ont eté realisés sur les couples de gènes uniques dont la structure intron-exon était validée experimentalement par la presence de transcrits. L'ensemble de ces observations, dont certaines sont indépendantes de la conservation de séquences, suggère très fortement que ces gènes uniques conservés dans *A. thaliana* et *O. sativa* sont *stricto sensu* des orthologues.

Les particularités des gènes uniques par rapport à l'ensemble des gènes nucléaires m'ont conduit à m'intéresser à leur conservation dans les génomes d'autres représentants du règne végétal. Malgré la stringence de notre sélection des gènes uniques, 192 gènes sont conservés uniques dans chacun des génomes d'*A. thaliana*, d'*O. sativa*, de la mousse *P. patens* et de l'algue unicellulaire *O. lucimarinus*. A titre comparatif, une conservation théorique aléatoire des gènes uniques entre *O. lucimarinus* et *A. thaliana* n'aboutit à la conservation apparente que d'environ 8-9 gènes. Ces gènes uniques hautement conservés sont majoritairement de fonction biochimique inconnue. Ils sont restés uniques malgré les divers évènements de duplication, globale, locale et segmentale, qui ont eu lieu depuis la spéciation des algues vertes il y a plus de 700 millions d'années. Une forte pression de sélection a donc maintenue unique un nombre important de gènes. Le profil phylogénétique des protéines codées par ces gènes uniques et l'étude de leur localisation subcellulaire prédite indique qu'environ 20% de ces gènes pourraient être d'origine bactérienne ou provenir d'un transfert d'ADN du chloroplaste au noyau. Le maintien de la stœchiométrie entre les produits de ces gènes et les protéines avec lesquelles ils formeraient des complexes fonctionnels dans les chloroplastes pourrait expliquer leur maintien à l'état unique dans le génome nucléaire après leur transfert.

De façon inattendue, les gènes uniques spécifiques soit d'*A.thaliana* soit du *O.sativa* présentent des caractéristiques communes. Ces gènes sont différents des gènes uniques conservés puisque, en moyenne, ils sont particulièrement courts, pauvres en introns et codent fréquemment pour des protéines dirigées vers le réticulum endoplasmique. Seulement 25% de ces gènes sont associés à une annotation GO et, parmi ceux-ci, se trouvent différentes classes de peptides auxquels il est attribué un rôle hormonal. Ces courts peptides proviennent de la maturation post-traductionnelle de pro-peptides qui, dans une famille fonctionnelle donnée comme les CLE, n'ont en commun qu'un trop petit nombre d'acides aminés pour être reconnu formellement comme homologues. Ces gènes uniques spécifiques d'une espèce semblent donc bien constituer une classe particulière de gènes dont la perte d'une copie après un événement de duplication serait favorisée soit par délétion soit par une évolution rapide de la séquence associée à la spéciation.

Les promoteurs des gènes uniques conservés contiennent un petit nombre de motifs sur-représentés

A l'issue de cette étude j'avais donc définis trois classes de gènes uniques présents chez *A.thaliana* et *O.sativa*. Les gènes uniques spécifiques d'une espèce, les couples de gènes orthologues uniques dans les deux espèces et des groupes de gènes comprenant un gène unique dans une espèce et différents homologues dans l'autre espèce. La composition en séquences régulatrices potentielles de leurs promoteurs a alors été analysée. J'ai utilisée une approche basée sur la notion d'empreintes phylogénétiques pour essayer de mettre en évidence la présence préférentielle de certaines séquences dans les couples de gènes orthologues uniques dans *A.thaliana* et *O.sativa*. Seulement les gènes pour lesquels le site d'initiation de transcription est expérimentalement défini ont été utilisés, c'est-à-dire un peu plus de la moitié des gènes dans les deux espèces. Un large panel de méthodes disponibles a été testé. MotifSampler s'est révélé être la méthode répondant le mieux aux conditions fixées: outils utilisable en local, facilité pour l'analyse automatique des résultats, rapidité et reconnaissance des séquences introduites dans le jeu d'essai.

J'ai également développé une méthode, DECOMO (pour DEscription of COnserved MOtifs), qui permet une description exhaustive des motifs conservés dans les deux promoteurs d'une paire de gènes uniques conservés. Pour diminuer le nombre de faux positifs possibles, c'est-à-dire augmenter la probabilité pour un motif d'être un motif fonctionnel, j'ai développé et appliqué un certain nombre de filtres : taille des motifs, ordre des motifs successifs, présence des motifs dans la même région des deux promoteurs et fréquences exceptionnelles. Ni MotifSampler ni DECOMO n'ont pu mettre en évidence une augmentation du nombre de motifs présents dans les promoteurs

des gènes uniques conservés chez *A. thaliana* et *O. sativa* en comparaison à des paires aléatoires de promoteurs. Des essais avec des couples de plantes phylogénétiquement plus proches, *A.thaliana* et radis ou peuplier et vigne, montrent un plus grand nombre de motifs en commun entre les gènes uniques conservés. Cependant, cet effet pourrait être du à la plus grande conservation globale entre les séquences promotrices et non pas la conservation de motifs fonctionnels *stricto sensu*.

J'ai alors introduit une méthode de prédiction *ab initio* de motifs pouvant potentiellement être des motifs de fixation de facteurs de transcription. La méthode utilisée est basée sur la localisation préférentielle de motifs dans les promoteurs. Elle a été implémentée au laboratoire par V. Bernard dans le cadre de sa thèse (actuellement en cours). Il y a 937 motifs qui forment un pic de fréquence à la même position dans les promoteurs d'*A. thaliana* et d'*O. sativa*.

La comparaison des différents groupes de gènes par cette méthode confirme que :

(i) il existe des motifs sur-représentés dans l'ensemble des gènes nucléaires et montre que :

(ii) l'on détecte moins de motifs dans les promoteurs des gènes uniques non conservés,

(iii) les promoteurs des gènes uniques conservés ont une sous représentation en boîte TATA,

(iv) il y a 13 motifs sur-représentés communs aux gènes uniques conservés entre *A. thaliana* et *O. sativa*,

(v) 25 motifs sur-représentés sont strictement spécifiques des gènes uniques.

Conclusions

Avant ce travail, un petit nombre d'études avaient été consacrées aux gènes uniques spécifiques d'une espèce. L'originalité de ma thèse a été d'étudier en parallèle, d'une part, les caractéristiques structurales et fonctionnelles et, d'autre part, la conservation au cours de l'évolution des différentes classes de gènes uniques. Mes résultats démontrent que les gènes uniques conservés et les gènes uniques non conservés ne partagent pas les mêmes caractéristiques tant au niveau de la structure des gènes et de la taille de la protéine codée par ces gènes, qu'au niveau fonctionnel et qu'au niveau de la présence de motifs dans les promoteurs. Ainsi, d'une part, les gènes uniques spécifiques à une espèce sont préférentiellement des gènes codant pour des peptides excrétés et impliqués dans des fonctions de régulation pouvant être liées à des mécanismes de défense ou de différentiation cellulaire. Tandis que, d'autre part, les gènes uniques conservés ont des caractéristiques qui les associent aux gènes impliqués dans les fonctions du métabolisme de base des cellules qui évoluent lentement. Par ailleurs, mes observations indiquent que, généralement, ces gènes, quand ils sont dupliqués, retournent rapidement à l'état unique. En conséquence, le caractère unique et la relation d'orthologie vraie est maintenue suggérant un fort désavantage sélectif à l'existence simultanée de deux copies peu divergentes de ces gènes.

Perspectives

A la suite de ce travail il est possible de proposer un certain nombre de pistes pour de futures études : (i) L'analyse des promoteurs des différentes classes de gènes uniques par la méthode *ab initio* basée sur la contrainte topologique et sa conservation pourrait être étendue à la région 5'UTR connue pour contenir des séquences régulatrices de la transcription. (ii) L'utilisation des données transcriptome d'*A. thaliana* et d'*O.sativa* pour comparer l'expression des gènes uniques conservés mais également pour explorer la nature des gènes uniques spécifiques d'une espèce et leur implication dans des voies de signalisation et de réponse à des facteurs biotiques. Cette approche constituerait la première étape dans la validation expérimentale de la fonction de 13 protéines riches en proline et glycine qui sont les produits de gènes uniques spécifiques d'une espèce et de 35 gènes végétaux conservés uniques mais aussi très largement conservés dans le monde vivant. (iii) Les raisons de l'existence des deux catégories de gènes uniques conservés, ceux conservés uniques et ceux montrant plusieurs homologues dans un des génomes, pourraient être recherchées dans l'ancienneté des duplications ou dans la force de la pression de sélection associée à la fonction de ces gènes.

CHAPTER VI

REFERENCES

VI. References

Adams, K. L., R. Cronn, R. Percifield and J. F. Wendel (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci U S A* **100**(8): 4649-54.

Aerts, S., G. Thijs, B. Coessens, M. Staes, Y. Moreau and B. De Moor (2003). Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res* **31**(6): 1753-64.

Ahuja, M. R. (1965). Genetic control of tumour formation in higher plants. *Q. Rev. Biol.* **40**: 329-340.

Aitkin, M. and D. B. Rubin (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society Serie B* **47**(1): 67-75.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. *J Mol Biol* **215**(3): 403-10.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-402.

Amores, A., A. Force, Y. L. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, J. Langeland, V. Prince, Y. L. Wang*, et al.* (1998). Zebrafish hox clusters and vertebrate genome evolution. *Science* **282**(5394): 1711-4.

Anisimova, M., J. P. Bielawski and Z. Yang (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* **18**(8): 1585-92.

Arisue, N., T. Hashimot, J. A. Lee, D. V. Moore, P. Gordon, C. W. Sensen, T. Gaasterland, M. Hasegawa and M. Muller (2002). The phylogenetic position of the pelobiont Mastigamoeba balamuthi based on sequences of rDNA and translation elongation factors EF-1alpha and EF-2. *J Eukaryot Microbiol* **49**(1): 1-10.

Arnone, M. I. and E. H. Davidson (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**(10): 1851-64.

Arnosti, D. N. and M. M. Kulkarni (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**(5): 890-8.

Aubourg, S., V. Brunaud, C. Bruyere, M. Cock, R. Cooke, A. Cottet, A. Couloux, P. Dehais, G. Deleage, A. Duclert*, et al.* (2005). GeneFarm, structural and functional annotation of Arabidopsis gene and protein families by a network of experts. *Nucleic Acids Res* **33**(Database issue): D641-6.

Aubourg, S., M. L. Martin-Magniette, V. Brunaud, L. Taconnat, F. Bitton, S. Balzergue, P. E. Jullien, M. Ingouff, V. Thareau, T. Schiex*, et al.* (2007). Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. *BMC Genomics* **8**: 401.

Aury, J. M., O. Jaillon, L. Duret, B. Noel, C. Jubin, B. M. Porcel, B. Segurens, V. Daubin, V. Anthouard, N. Aiach*, et al.* (2006). Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**(7116): 171-8.

Ayala, F. J. (1999). Molecular clock mirages. *Bioessays* **21**(1): 71-5.

Bailey, T. L. and C. Elkan (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28-36.

Basehoar, A. D., S. J. Zanton and B. F. Pugh (2004). Identification and distinct regulation of yeast TATA box-containing genes. *Cell* **116**(5): 699-709.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer*, et al.* (2004). The Pfam protein families database. *Nucleic Acids Res* **32**(Database issue): D138-41.

Beckwith, J. R. (1967). Regulation of the lac operon. Recent studies on the regulation of lactose metabolism in Escherichia coli support the operon model. *Science* **156**(3775): 597-604.

Bellora, N., D. Farre and M. M. Alba (2007). Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. *BMC Genomics* **8**: 459.

Benabdellah, K., E. Gonzalez-Rey and A. Gonzalez (2007). Alternative trans-splicing of the Trypanosoma cruzi LYT1 gene transcript results in compartmental and functional switch for the encoded protein. *Mol Microbiol* **65**(6): 1559-67.

Benoist, C., K. O'Hare, R. Breathnach and P. Chambon (1980). The ovalbumin gene-sequence of putative control regions. *Nucleic Acids Res* **8**(1): 127-42.

Berendzen, K. W., K. Stuber, K. Harter and D. Wanke (2006). Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves. *BMC Bioinformatics* **7**: 522.

Berezikov, E., V. Guryev and E. Cuppen (2005). CONREAL web server: identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Res* **33**(Web Server issue): W447-50.

Bernard, V., V. Brunaud, C. Serizet, M. L. Martin-Magniette, M. Caboche, S. Aubourg and A. Lecharny (2006). Sélection de motifs candidats pour la régulation des gènes chez Arabidopsis thaliana sur des critères topologiques. JOBIM. Bordeaux**:** 17-28.

Birchler, J. A., H. Yao and S. Chudalayandi (2007). Biological consequences of dosage dependent gene regulatory systems. *Biochim Biophys Acta* **1769**(5-6): 422-8.

Blair, J. E., P. Shah and S. B. Hedges (2005). Evolutionary sequence analysis of complete eukaryote genomes. *BMC Bioinformatics* **6**: 53.

Blanc, G. and K. H. Wolfe (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* **16**(7): 1679-91.

Blanchette, M., B. Schwikowski and M. Tompa (2002). Algorithms for phylogenetic footprinting. *J Comput Biol* **9**(2): 211-23.

Blanchette, M. and M. Tompa (2003). FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res* **31**(13): 3840-2.

Blomme, T., K. Vandepoele, S. De Bodt, C. Simillion, S. Maere and Y. Van de Peer (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7**(5): R43.

Boffelli, D., J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter and E. M. Rubin (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**(5611): 1391-4.

Boudet, N., S. Aubourg, C. Toffano-Nioche, M. Kreis and A. Lecharny (2001). Evolution of intron/exon structure of DEAD helicase family genes in Arabidopsis, Caenorhabditis, and Drosophila. *Genome Res* **11**(12): 2101-14.

Bowers, J. E., B. A. Chapman, J. Rong and A. H. Paterson (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**(6930): 433-8.

Brasset, E. and C. Vaury (2005). Insulators are fundamental components of the eukaryotic genomes. *Heredity* **94**(6): 571-6.

Bray, N., I. Dubchak and L. Pachter (2003). AVID: A global alignment program. *Genome Res* **13**(1): 97-102.

Bray, N. and L. Pachter (2004). MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* **14**(4): 693-9.

Brideau, N. J., H. A. Flores, J. Wang, S. Maheshwari, X. Wang and D. A. Barbash (2006). Two Dobzhansky-Muller genes interact to cause hybrid lethality in Drosophila. *Science* **314**(5803): 1292-5.

Brooks, A. R., B. P. Nagy, S. Taylor, W. S. Simonet, J. M. Taylor and B. Levy-Wilson (1994). Sequences containing the second-intron enhancer are essential for transcription of the human apolipoprotein B gene in the livers of transgenic mice. *Mol Cell Biol* **14**(4): 2243-56.

Bru, C., E. Courcelle, S. Carrere, Y. Beausse, S. Dalmar and D. Kahn (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res* **33**(Database issue): D212-5.

Brudno, M., C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow and S. Batzoglou (2003). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**(4): 721-31.

Bucher, P. and E. N. Trifonov (1988). CCAAT box revisited: bidirectionality, location and context. *J Biomol Struct Dyn* **5**(6): 1231-6.

Buratowski, S., S. Hahn, P. A. Sharp and L. Guarente (1988). Function of a yeast TATA element-binding protein in a mammalian transcription system. *Nature* **334**(6177): 37-42.

Burgess-Beusse, B., C. Farrell, M. Gaszner, M. Litt, V. Mutskov, F. Recillas-Targa, M. Simpson, A. West and G. Felsenfeld (2002). The insulation of genes from external enhancers and silencing chromatin. *Proc Natl Acad Sci U S A* **99 Suppl 4**: 16433-7.

Buttner, M. and K. B. Singh (1997). Arabidopsis thaliana ethylene-responsive element binding protein (AtEBP), an ethylene-inducible, GCC box DNA-binding protein interacts with an ocs element binding protein. *Proc Natl Acad Sci U S A* **94**(11): 5961-6.

Carmel, L., I. B. Rogozin, Y. I. Wolf and E. V. Koonin (2007). Evolutionarily conserved genes preferentially accumulate introns. *Genome Res* **17**(7): 1045-50.

Carmel, L., Y. I. Wolf, I. B. Rogozin and E. V. Koonin (2007). Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res* **17**(7): 1034-44.

Cavalier-Smith, T. and E. E. Chao (2003). Phylogeny and classification of phylum Cercozoa (Protozoa). *Protist* **154**(3-4): 341-58.

Clark, A. G. (2001). The search for meaning in noncoding DNA. *Genome Res* **11**(8): 1319-20.

Clauss, M. J. and T. Mitchell-Olds (2004). Functional divergence in tandemly duplicated Arabidopsis thaliana trypsin inhibitor genes. *Genetics* **166**(3): 1419-36.

Cook, P. R. (2003). Nongenic transcription, gene regulation and action at a distance. *J Cell Sci* **116**(Pt 22): 4483-91.

Cooper, G. M. and R. E. Hausman (2007). The cell : a molecular approach. Washington, DC , Sunderland, MA, ASM Press, Sinauer Associates.

Coyne, J. A. and H. A. Orr (2004). Speciation. Sunderland, MA, Sinauer Associates.

Cronn, R. C., R. L. Small, T. Haselkorn and J. F. Wendel (2002). Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *Am J Bot* **84**: 707-725.

Chantret, N., J. Salse, F. Sabot, S. Rahman, A. Bellec, B. Laubin, I. Dubois, C. Dossat, P. Sourdille, P. Joudrier*, et al.* (2005). Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *Plant Cell* **17**(4): 1033-45.

Chapman, B. A., J. E. Bowers, F. A. Feltus and A. H. Paterson (2006). Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. *Proc Natl Acad Sci U S A* **103**(8): 2730-5.

Chaw, S. M., C. C. Chang, H. L. Chen and W. H. Li (2004). Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J Mol Evol* **58**(4): 424-41.

Chenna, R., H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins and J. D. Thompson (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**(13): 3497-500.

Chothia, C., J. Gough, C. Vogel and S. A. Teichmann (2003). Evolution of the protein repertoire. *Science* **300**(5626): 1701-3.

Dehal, P. and J. L. Boore (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**(10): e314.

Dempster, A., N. Laird and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Serie B* **39**(1): 1-38.

Desloire, S., H. Gherbi, W. Laloui, S. Marhadour, V. Clouet, L. Cattolico, C. Falentin, S. Giancola, M. Renard, F. Budar*, et al.* (2003). Identification of the fertility restoration locus, Rfo, in radish, as a member of the pentatricopeptide-repeat protein family. *EMBO Rep* **4**(6): 588-94.

Dietrich, F. S., S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pohlmann, P. Luedi, S. Choi*, et al.* (2004). The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* **304**(5668): 304-7.

Domazet-Loso, T. and D. Tautz (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome Res* **13**(10): 2213-9.

Drummond, D. A., A. Raval and C. O. Wilke (2006). A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* **23**(2): 327-37.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuveglise, E. Talla*, et al.* (2004). Genome evolution in yeasts. *Nature* **430**(6995): 35-44.

Duret, L. and P. Bucher (1997). Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* **7**(3): 399-406.

Dynlacht, B. D., T. Hoey and R. Tjian (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. *Cell* **66**(3): 563-76.

Ekker, M., J. Wegner, M. A. Akimenko and M. Westerfield (1992). Coordinate embryonic expression of three zebrafish engrailed genes. *Development* **116**(4): 1001-10.

Ermolaeva, M. D., M. Wu, J. A. Eisen and S. L. Salzberg (2003). The age of the Arabidopsis thaliana genome duplication. *Plant Mol Biol* **51**(6): 859-66.

Escriva, H., S. Bertrand, P. Germain, M. Robinson-Rechavi, M. Umbhauer, J. Cartry, M. Duffraisse, L. Holland, H. Gronemeyer and V. Laudet (2006). Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genet* **2**(7): e102.

Farrokhi, N., J. P. Whitelegge and J. A. Brusslan (2008). Plant peptides and peptidomics. *Plant Biotechnol J* **6**(2): 105-34.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool* **19**(2): 99-113.

Flavell, R. B. (1994). Inactivation of gene expression in plants as a consequence of specific sequence duplication. *Proc Natl Acad Sci U S A* **91**(9): 3490-6.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan and J. Postlethwait (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**(4): 1531-45.

Frech, K., G. Herrmann and T. Werner (1993). Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res* **21**(7): 1655-64.

Frith, M. C., A. R. Forrest, E. Nourbakhsh, K. C. Pang, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, T. L. Bailey and S. M. Grimmond (2006). The abundance of short proteins in the mammalian proteome. *PLoS Genet* **2**(4): e52.

Fujimori, S., T. Washio, K. Higo, Y. Ohtomo, K. Murakami, K. Matsubara, J. Kawai, P. Carninci, Y. Hayashizaki, S. Kikuchi*, et al.* (2003). A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *FEBS Lett* **554**(1-2): 17-22.

Fukuchi, S. and K. Nishikawa (2003). Estimation of the Number of Orphan Genes in the Genome Sequences. *Genome Informatics* **14**: 468-469.

Gagnot, S., J. P. Tamby, M. L. Martin-Magniette, F. Bitton, L. Taconnat, S. Balzergue, S. Aubourg, J. P. Renou, A. Lecharny and V. Brunaud (2008). CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Res* **36**(Database issue): D986-90.

Gaillardin, C., G. Duchateau-Nguyen, F. Tekaia, B. Llorente, S. Casaregola, C. Toffano-Nioche, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara*, et al.* (2000). Genomic exploration of the hemiascomycetous yeasts: 21. Comparative functional classification of genes. *FEBS Lett* **487**(1): 134-49.

Galindo, M. I., J. I. Pueyo, S. Fouix, S. A. Bishop and J. P. Couso (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**(5): e106.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6): 721-741.

Gershenzon, N. I. and I. P. Ioshikhes (2005). Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* **21**(8): 1295-300.

Glemet, E. and J. J. Codani (1997). LASSAP, a LArge Scale Sequence compArison Package. *Comput Appl Biosci* **13**(2): 137-43.

Gogarten, J. P., W. F. Doolittle and J. G. Lawrence (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**(12): 2226-38.

Gollery, M., J. Harper, J. Cushman, T. Mittler, T. Girke, J. K. Zhu, J. Bailey-Serres and R. Mittler (2006). What makes species unique? The contribution of proteins with obscure features. *Genome Biol* **7**(7): R57.

Goodyer, C. G., G. Zogopoulos, G. Schwartzbauer, H. Zheng, G. N. Hendy and R. K. Menon (2001). Organization and evolution of the human growth hormone receptor gene 5'-flanking region. *Endocrinology* **142**(5): 1923-34.

Grad, Y. H., F. P. Roth, M. S. Halfon and G. M. Church (2004). Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura. *Bioinformatics* **20**(16): 2738-50.

Guldener, U., M. Munsterkotter, G. Kastenmuller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. Garcia-Martinez, J. E. Perez-Ortin*, et al.* (2005). CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res* **33**(Database issue): D364-8.

Gutierrez, R. A., P. J. Green, K. Keegstra and J. B. Ohlrogge (2004). Phylogenetic profiling of the Arabidopsis thaliana proteome: what proteins distinguish plants from other organisms? *Genome Biol* **5**(8): R53.

Guyot, R. and B. Keller (2004). Ancestral genome duplication in rice. *Genome* **47**(3): 610-4.

Haerty, W. and R. S. Singh (2006). Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of Drosophila. *Mol Biol Evol* **23**(9): 1707-14.

Hanada, K., X. Zhang, J. O. Borevitz, W. H. Li and S. H. Shiu (2007). A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Res* **17**(5): 632-40.

Harrison, P. M., N. Carriero, Y. Liu and M. Gerstein (2003). A "polyORFomic" analysis of prokaryote genomes using disabled-homology filtering reveals conserved but undiscovered short ORFs. *J Mol Biol* **333**(5): 885-92.

Hedges, S. B., J. E. Blair, M. L. Venturi and J. L. Shoe (2004). A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol* **4**: 2.

Hertz, G. Z. and G. D. Stormo (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**(7-8): 563-77.

Hindemitt, T. and K. F. Mayer (2005). CREDO: a web-based tool for computational detection of conserved sequence motifs in noncoding sequences. *Bioinformatics* **21**(23): 4304-6.

Hoebeke, M. and S. Schbath (2006). R'MES: Finding Exceptional Motifs.

Hoey, T., B. D. Dynlacht, M. G. Peterson, B. F. Pugh and R. Tjian (1990). Isolation and characterization of the Drosophila gene encoding the TATA box binding protein, TFIID. *Cell* **61**(7): 1179-86.

Holland, P. W. (1997). Vertebrate evolution: something fishy about Hox genes. *Curr Biol* **7**(9): R570-2.

Hong, R. L., L. Hamaguchi, M. A. Busch and D. Weigel (2003). Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. *Plant Cell* **15**(6): 1296-309.

Horan, K., C. Jang, J. Bailey-Serres, R. Mittler, C. Shelton, J. F. Harper, J. K. Zhu, J. C. Cushman, M. Gollery and T. Girke (2008). Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol* **147**(1): 41-57.

Hudson, M. E. and P. H. Quail (2003). Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. *Plant Physiol* **133**(4): 1605-16.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**(1346): 119-24.

Hughes, J. D., P. W. Estep, S. Tavazoie and G. M. Church (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**(5): 1205-14.

Hughes, T., D. Ekman, H. Ardawatia, A. Elofsson and D. A. Liberles (2007). Evaluating dosage compensation as a cause of duplicate gene retention in Paramecium tetraurelia. *Genome Biol* **8**(5): 213.

Jacob, F., D. Perrin, C. Sanchez and J. Monod (1960). [Operon: a group of genes with the expression coordinated by an operator.]. *C R Hebd Seances Acad Sci* **250**: 1727-9.

Jaillon, O., J. M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin*, et al.* (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**(7161): 463-7.

Jensen, L. J. and S. Knudsen (2000). Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics* **16**(4): 326-33.
JGI. JGI. from http://genome.jgi-psf.org/.
Jones, A. M., J. Chory, J. L. Dangl, M. Estelle, S. E. Jacobsen, E. M. Meyerowitz, M. Nordborg and D. Weigel (2008). The impact of Arabidopsis on human health: diversifying our portfolio. *Cell* **133**(6): 939-43.
Joshi, M. G. (1972). Occurrence of genetic tumours in Triticum interspecies hybrids. *Theor. Appl. Genet.* **42**: 227-228.
Joyner, A. L. and G. R. Martin (1987). En-1 and En-2, two mouse genes with sequence homology to the Drosophila engrailed gene: expression during embryogenesis. *Genes Dev* **1**(1): 29-38.
Kaplinsky, N. J., D. M. Braun, J. Penterman, S. A. Goff and M. Freeling (2002). Utility and distribution of conserved noncoding sequences in the grasses. *Proc Natl Acad Sci U S A* **99**(9): 6147-51.
Kastenmayer, J. P., L. Ni, A. Chu, L. E. Kitchen, W. C. Au, H. Yang, C. D. Carter, D. Wheeler, R. W. Davis, J. D. Boeke*, et al.* (2006). Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res* **16**(3): 365-73.
Kellis, M., B. W. Birren and E. S. Lander (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**(6983): 617-24.
Kielbasa, S. M., J. O. Korbel, D. Beule, J. Schuchhardt and H. Herzel (2001). Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics* **17**(11): 1019-26.
Kim, T. H., L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green and B. Ren (2005). A high-resolution map of active promoters in the human genome. *Nature* **436**(7052): 876-80.
Knowles, D. G. and A. McLysaght (2006). High rate of recent intron gain and loss in simultaneously duplicated Arabidopsis genes. *Mol Biol Evol* **23**(8): 1548-57.
Koch, M., B. Haubold and T. Mitchell-Olds (2001). Molecular systematics of the Brassicaceae: evidence from coding plastidic matK and nuclear Chs sequences. *Am J Bot* **88**(3): 534-544.
Koch, M. A., B. Haubold and T. Mitchell-Olds (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). *Mol Biol Evol* **17**(10): 1483-98.
Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf and E. V. Koonin (2002). Selection in the evolution of gene duplications. *Genome Biol* **3**(2): RESEARCH0008.
Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**: 309-38.
Koszul, R., S. Caburet, B. Dujon and G. Fischer (2004). Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *EMBO J* **23**(1): 234-43.
Krasowski, M. D., K. Yasuda, L. R. Hagey and E. G. Schuetz (2005). Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). *Nucl Recept* **3**: 2.
Krylov, D. M., Y. I. Wolf, I. B. Rogozin and E. V. Koonin (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* **13**(10): 2229-35.
Ku, H. M., T. Vision, J. Liu and S. D. Tanksley (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. *Proc Natl Acad Sci U S A* **97**(16): 9121-6.
Kutach, A. K. and J. T. Kadonaga (2000). The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol Cell Biol* **20**(13): 4754-64.
Labedan, B. and M. Riley (1995). Gene products of Escherichia coli: sequence comparisons and common ancestries. *Mol Biol Evol* **12**(6): 980-7.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh*, et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.
Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald and J. C. Wootton (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**(5131): 208-14.
Lawton-Rauh, A. (2003). Evolutionary dynamics of duplicated genes in plants. *Mol Phylogenet Evol* **29**(3): 396-409.
Le Bouder-Langevin, S., I. Capron-Montaland, R. De Rosa and B. Labedan (2002). A strategy to retrieve the whole set of protein modules in microbial proteomes. *Genome Res* **12**(12): 1961-73.
Lease, K. A. and J. C. Walker (2006). The Arabidopsis unannotated secreted peptide database, a resource for plant peptidomics. *Plant Physiol* **142**(3): 831-8.
Lee, T. I. and R. A. Young (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34**: 77-137.
Leung, J. Y., F. E. McKenzie, A. M. Uglialoro, P. O. Flores-Villanueva, B. C. Sorkin, E. J. Yunis, D. L. Hartl and A. E. Goldfeld (2000). Identification of phylogenetic footprints in primate tumor necrosis factor-alpha promoters. *Proc Natl Acad Sci U S A* **97**(12): 6614-8.
Li, X. and M. Noll (1994). Evolution of distinct developmental functions of three Drosophila genes by acquisition of different cis-regulatory regions. *Nature* **367**(6458): 83-7.
Li, X., S. Zhong and W. H. Wong (2005). Reliable prediction of transcription factor binding sites by phylogenetic verification. *Proc Natl Acad Sci U S A* **102**(47): 16945-50.

Liang, P., B. Labedan and M. Riley (2002). Physiological genomics of Escherichia coli protein families. *Physiol Genomics* **9**(1): 15-26.

Lifton, R. P., M. L. Goldberg, R. W. Karp and D. S. Hogness (1978). The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harb Symp Quant Biol* **42 Pt 2**: 1047-51.

Lindsey, K., S. Casson and P. Chilley (2002). Peptides: new signalling molecules in plants. *Trends Plant Sci* **7**(2): 78-83.

Linial, M. (2003). How incorrect annotations evolve--the case of short ORFs. *Trends Biotechnol* **21**(7): 298-300.

Liu, J. S. and C. E. Lawrence (1999). Bayesian inference on biopolymer models. *Bioinformatics* **15**(1): 38-52.

Liu, Y., Q. Zhu and N. Zhu (2008). Recent duplication and positive selection of the GAGE gene family. *Genetica* **133**(1): 31-5.

Lobo-Menendez, F., L. H. Bowman and M. J. Dewey (2004). Inverted GCG/CGC trinucleotide microsatellites in the 5'-region of Mus IDS mRNA: recurrent induction of aberrant reverse transcripts. *Mol Biol Rep* **31**(2): 107-12.

Lynch, M. and J. S. Conery (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**(5494): 1151-5.

Lynch, M. and J. S. Conery (2003). The evolutionary demography of duplicate genes. *J Struct Funct Genomics* **3**(1-4): 35-44.

Lynch, M. and A. Force (2000). Gene duplication and the origin of interspecific genomic incompatibility. *The American Naturalist* **156**: 590-605.

Lynch, M., M. O'Hely, B. Walsh and A. Force (2001). The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**(4): 1789-804.

Lynch, M. and B. Walsh (1998). Genetics and Analysis of Quantitative Traits. Sunderland, MA, Sinauer Associates, Inc.

Llorente, B., P. Durrens, A. Malpertuy, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, E. Bon, P. Brottier, S. Casaregola*, et al.* (2000). Genomic exploration of the hemiascomycetous yeasts: 20. Evolution of gene redundancy compared to Saccharomyces cerevisiae. *FEBS Lett* **487**(1): 122-33.

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper and Y. Van de Peer (2005). Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A* **102**(15): 5454-9.

Makino, T., Y. Suzuki and T. Gojobori (2006). Differential evolutionary rates of duplicated genes in protein interaction network. *Gene* **385**: 57-63.

Matys, V., E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis*, et al.* (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**(1): 374-8.

McClintock, J. M., R. Carlson, D. M. Mann and V. E. Prince (2001). Consequences of Hox gene duplication in the vertebrates: an investigation of the zebrafish Hox paralogue group 1 genes. *Development* **128**(13): 2471-84.

McCue, L., W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire and C. E. Lawrence (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* **29**(3): 774-82.

Mohseni-Zadeh, S., A. Louis, P. Brezellec and J. L. Risler (2004). PHYTOPROT: a database of clusters of plant proteins. *Nucleic Acids Res* **32**(Database issue): D351-3.

Molina, C. and E. Grotewold (2005). Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* **6**(1): 25.

Morgenstern, B., A. Dress and T. Werner (1996). Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc Natl Acad Sci U S A* **93**(22): 12098-103.

Morgenstern, B., K. Frech, A. Dress and T. Werner (1998). DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics* **14**(3): 290-4.

Moskal, W. A., Jr., H. C. Wu, B. A. Underwood, W. Wang, C. D. Town and Y. Xiao (2007). Experimental validation of novel genes predicted in the un-annotated regions of the Arabidopsis genome. *BMC Genomics* **8**: 18.

Needleman, S. B. and C. D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**(3): 443-53.

Nei, M. and T. Gojobori (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**(5): 418-26.

Nekrutenko, A., K. D. Makova and W. H. Li (2002). The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res* **12**(1): 198-202.

Neuwald, A. F., J. S. Liu and C. E. Lawrence (1995). Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci* **4**(8): 1618-32.

Nikolaev, S. I., C. Berney, J. F. Fahrni, I. Bolivar, S. Polet, A. P. Mylnikov, V. V. Aleshin, N. B. Petrov and J. Pawlowski (2004). The twilight of Heliozoa and rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. *Proc Natl Acad Sci U S A* **101**(21): 8066-71.

Ogbourne, S. and T. M. Antalis (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochim J* **331**(Pt 1): 1-14.

Ohme-Takagi, M. and H. Shinshi (1995). Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element. *Plant Cell* **7**(2): 173-82.

Ohno, S. (1970). Evolution by Gene Duplication. Berlin-Heidelberg-New York, Springer-Verlag.
Orr, H. A. (1996). Dobzhansky, Bateson, and the genetics of speciation. *Genetics* **144**(4): 1331-5.
Orr, H. A. and S. Irving (2001). Complex epistasis and the genetic basis of hybrid sterility in the Drosophila pseudoobscura Bogota-USA hybridization. *Genetics* **158**(3): 1089-100.
Pal, C., B. Papp and L. D. Hurst (2001). Highly expressed genes in yeast evolve slowly. *Genetics* **158**(2): 927-31.
Palenik, B., J. Grimwood, A. Aerts, P. Rouze, A. Salamov, N. Putnam, C. Dupont, R. Jorgensen, E. Derelle, S. Rombauts*, et al.* (2007). The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A* **104**(18): 7705-10.
Pan, Y., S. Phan, A. F. Famili, A. E. G. Lenfering, M. L. Jaramillo and M. O'connor-Mccourt (2008). Do Orthologous Genes Have Orthologous Promoters? ISMB. Toronto.
Papp, B., C. Pal and L. D. Hurst (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**(6945): 194-7.
Paterson, A. H., J. E. Bowers and B. A. Chapman (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A* **101**(26): 9903-8.
Patikoglou, G. A., J. L. Kim, L. Sun, S. H. Yang, T. Kodadek and S. K. Burley (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev* **13**(24): 3217-30.
Pebusque, M. J., F. Coulier, D. Birnbaum and P. Pontarotti (1998). Ancient large-scale genome duplications: phylogenetic and linkage analyses shed light on chordate genome evolution. *Mol Biol Evol* **15**(9): 1145-59.
Philippe, H. and J. Laurent (1998). How good are deep phylogenetic trees? *Curr Opin Genet Dev* **8**(6): 616-23.
Philippe, H., P. Lopez, H. Brinkmann, K. Budin, A. Germot, J. Laurent, D. Moreira, M. Muller and H. Le Guyader (2000). Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc Biol Sci* **267**(1449): 1213-21.
Phillips, L. L. and R. K. Reid (1975). Interspecific Incompatibility In Gossypium. II. Light and electron microscopic studies of cell necrosis and tumorigenesis in hybrids of G. klotzschianum. *Am J Bot* **62**: 790-796.
Pirkkala, L., P. Nykanen and L. Sistonen (2001). Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. *FASEB J* **15**(7): 1118-31.
Pollard, D. A., C. M. Bergman, J. Stoye, S. E. Celniker and M. B. Eisen (2004). Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**: 6.
Pollard, D. A., C. M. Bergman, J. Stoye, S. E. Celniker and M. B. Eisen (2004). Correction: Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**: 73.
Postlethwait, J., A. Amores, W. Cresko, A. Singer and Y. L. Yan (2004). Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet* **20**(10): 481-90.
Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc Natl Acad Sci U S A* **72**(3): 784-8.
Reinberg, D., G. Orphanides, R. Ebright, S. Akoulitchev, J. Carcamo, H. Cho, P. Cortes, R. Drapkin, O. Flores, I. Ha*, et al.* (1998). The RNA polymerase II general transcription factors: past, present, and future. *Cold Spring Harb Symp Quant Biol* **63**: 83-103.
Rensing, S. A., D. Lang, A. D. Zimmer, A. Terry, A. Salamov, H. Shapiro, T. Nishiyama, P. F. Perroud, E. A. Lindquist, Y. Kamisugi*, et al.* (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**(5859): 64-9.
Rice Chromosomes 11 and 12 Sequencing Consortia (2005). The sequence of rice chromosomes 11 and 12, rich in disease resistance genes and recent gene duplications. *BMC Biol* **3**: 20.
Riley, M. and B. Labedan (1997). Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module. *J Mol Biol* **268**(5): 857-68.
Rodriguez-Ezpeleta, N., H. Brinkmann, S. C. Burey, B. Roure, G. Burger, W. Loffelhardt, H. J. Bohnert, H. Philippe and B. F. Lang (2005). Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr Biol* **15**(14): 1325-30.
Roest Crollius, H., O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau, C. Fischer, C. Fizames, P. Wincker, P. Brottier, F. Quetier*, et al.* (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat Genet* **25**(2): 235-8.
Roth, F. P., J. D. Hughes, P. W. Estep and G. M. Church (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**(10): 939-45.
Roy, S. W. and D. Penny (2007). Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of O. sativa and A. thaliana. *Mol Biol Evol* **24**(1): 171-81.
Rubin, G. M., M. D. Yandell, J. R. Wortman, G. L. Gabor Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann*, et al.* (2000). Comparative genomics of the eukaryotes. *Science* **287**(5461): 2204-15.
Sakai, H. and T. Itoh (2008). Reductive genome evolution during the course of Asian rice cultivation. SMBE. Barcelona.

Salse, J., S. Bolot, M. Throude, V. Jouffe, B. Piegu, U. M. Quraishi, T. Calcagno, R. Cooke, M. Delseny and C. Feuillet (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**(1): 11-24.

Samson, F., V. Brunaud, S. Duchene, Y. De Oliveira, M. Caboche, A. Lecharny and S. Aubourg (2004). FLAGdb++: a database for the functional analysis of the Arabidopsis genome. *Nucleic Acids Res* **32**(Database issue): D347-50.

Sandelin, A., W. Alkema, P. Engstrom, W. W. Wasserman and B. Lenhard (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**(Database issue): D91-4.

Scannell, D. R., K. P. Byrne, J. L. Gordon, S. Wong and K. H. Wolfe (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440**(7082): 341-5.

Scannell, D. R., A. C. Frank, G. C. Conant, K. P. Byrne, M. Woolfit and K. H. Wolfe (2007). Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A* **104**(20): 8397-402.

Schnable, P. S. and R. P. Wise (1998). The molecular basis of cytoplasmic male sterility. *Trends in Plant Science* **3**: 175-180.

Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler and W. Miller (2003). Human-mouse alignments with BLASTZ. *Genome Res* **13**(1): 103-7.

Semon, M. and K. H. Wolfe (2007). Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet* **23**(3): 108-12.

Seoighe, C. and C. Gehring (2004). Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. *Trends Genet* **20**(10): 461-4.

Seoighe, C. and K. H. Wolfe (1999). Updated map of duplicated regions in the yeast genome. *Gene* **238**(1): 253-61.

Shinshi, H., S. Usami and M. Ohme-Takagi (1995). Identification of an ethylene-responsive region in the promoter of a tobacco class I chitinase gene. *Plant Mol Biol* **27**(5): 923-32.

Simillion, C., K. Vandepoele, M. C. Van Montagu, M. Zabeau and Y. Van de Peer (2002). The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **99**(21): 13627-32.

Simmen, M. W., S. Leitgeb, V. H. Clark, S. J. Jones and A. Bird (1998). Gene number in an invertebrate chordate, Ciona intestinalis. *Proc Natl Acad Sci U S A* **95**(8): 4437-40.

Skovgaard, M., L. J. Jensen, S. Brunak, D. Ussery and A. Krogh (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet* **17**(8): 425-8.

Smale, S. T. and D. Baltimore (1989). The "initiator" as a transcription control element. *Cell* **57**(1): 103-13.

Smith, T. F. and M. S. Waterman (1981). Comparison of biosequences. *Advances in Applied Mathematics* **2**: 482-489.

Snyder, M. and M. Gerstein (2003). Genomics. Defining genes in the genomics era. *Science* **300**(5617): 258-60.

Sogin, M. L. (1991). Early evolution and the origin of eukaryotes. *Curr Opin Genet Dev* **1**(4): 457-63.

Sonnhammer, E. L. and D. Kahn (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci* **3**(3): 482-92.

Sonnhammer, E. L. and E. V. Koonin (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* **18**(12): 619-20.

Spring, J. (1997). Vertebrate evolution by interspecific hybridisation--are we polyploid? *FEBS Lett* **400**(1): 2-8.

Sterck, L., S. Rombauts, K. Vandepoele, P. Rouze and Y. Van de Peer (2007). How many genes are there in plants (... and why are they there)? *Curr Opin Plant Biol* **10**(2): 199-203.

Stoye, J., D. Evers and F. Meyer (1998). Rose: generating sequence families. *Bioinformatics* **14**(2): 157-63.

Susko, E., Y. Inagaki and A. J. Roger (2004). On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. *Mol Biol Evol* **21**(9): 1629-42.

Swarbreck, D., C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz*, et al.* (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* **36**(Database issue): D1009-14.

Syvanen, M. (1985). Cross-species gene transfer; implications for a new theory of evolution. *J Theor Biol* **112**(2): 333-43.

Tagle, D. A., B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess and R. T. Jones (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* **203**(2): 439-55.

TAIR. TAIR. from http://www.arabidopsis.org/.

Tanese, N., B. F. Pugh and R. Tjian (1991). Coactivators for a proline-rich activator purified from the multisubunit human TFIID complex. *Genes Dev* **5**(12A): 2212-24.

Tatusov, R. L., E. V. Koonin and D. J. Lipman (1997). A genomic perspective on protein families. *Science* **278**(5338): 631-7.

Taylor, J. S. and J. Raes (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**: 615-43.

Taylor, J. S., Y. Van de Peer and A. Meyer (2001). Genome duplication, divergent resolution and speciation. *Trends Genet* **17**(6): 299-301.

Tekaia, F. and B. Dujon (1999). Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J Mol Evol* **49**(5): 591-600.

Termier, M. and A. Kalogeropoulos (1996). Discrimination between fortuitous and biologically constrained open reading frames in DNA sequences of Saccharomyces cerevisiae. *Yeast* **12**(4): 369-84.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**(6814): 796-815.

Thijs, G., M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze and Y. Moreau (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**(12): 1113-22.

Thomas-Chollier, M., O. Sand, J. V. Turatsinze, R. Janky, M. Defrance, E. Vervisch, S. Brohee and J. van Helden (2008). RSAT: regulatory sequence analysis tools. *Nucleic Acids Res* **36**(Web Server issue): W119-27.

TIGR. TIGR. from http://www.tigr.org.

Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proc Int Conf Intell Syst Mol Biol*: 262-71.

Tompa, M., N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent*, et al.* (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **23**(1): 137-44.

Tremousaygue, D., L. Garnier, C. Bardet, P. Dabos, C. Herve and B. Lescure (2003). Internal telomeric repeats and 'TCP domain' protein-binding sites co-operate to regulate gene expression in Arabidopsis thaliana cycling cells. *Plant J* **33**(6): 957-66.

Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov*, et al.* (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**(5793): 1596-604.

van Helden, J. (2003). Regulatory sequence analysis tools. *Nucleic Acids Res* **31**(13): 3593-6.

van Helden, J., B. Andre and J. Collado-Vides (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* **281**(5): 827-42.

van Helden, J., A. F. Rios and J. Collado-Vides (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* **28**(8): 1808-18.

Vandepoele, K. and Y. Van de Peer (2005). Exploring the plant transcriptome through phylogenetic profiling. *Plant Physiol* **137**(1): 31-42.

Vanet, A., L. Marsan, A. Labigne and M. F. Sagot (2000). Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals. *J Mol Biol* **297**(2): 335-53.

Wagner, A. (2002). Asymmetric functional divergence of duplicate genes in yeast. *Mol Biol Evol* **19**(10): 1760-8.

Wagner, A. (2005). Energy constraints on the evolution of gene expression. *Mol Biol Evol* **22**(6): 1365-74.

Wainright, P. O., G. Hinkle, M. L. Sogin and S. K. Stickel (1993). Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science* **260**(5106): 340-2.

Walther, D., R. Brunnemann and J. Selbig (2007). The regulatory code for transcriptional response diversity and its relation to genome structural properties in A. thaliana. *PLoS Genet* **3**(2): e11.

Wang, G. and W. Zhang (2006). A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements. *Genome Biol* **7**(6): R49.

Wang, Q. F., S. Prabhakar, S. Chanan, J. F. Cheng, E. M. Rubin and D. Boffelli (2007). Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. *Genome Biol* **8**(1): R1.

Wapinski, I., A. Pfeffer, N. Friedman and A. Regev (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**(7158): 54-61.

Welch, J. J. (2004). Accumulating Dobzhansky-Muller incompatibilities: reconciling theory and data. *Evolution* **58**(6): 1145-56.

Welch, J. J., E. Fontanillas and L. Bromham (2005). Molecular dates for the "cambrian explosion": the influence of prior assumptions. *Syst Biol* **54**(4): 672-8.

Werth, C. R. and M. D. Windham (1991). A model for divergent allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. *The American Naturalist* **137**: 515-526.

Wilkins, A. S., Ed. (2002). The evolution of developmental pathways. Sunderland, MA, Sinaur Associates, Inc.

Wolfe, K. H., M. Gouy, Y. W. Yang, P. M. Sharp and W. H. Li (1989). Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc Natl Acad Sci U S A* **86**(16): 6201-5.

Wolfe, K. H. and D. C. Shields (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**(6634): 708-13.

Wong, J. M. and E. Bateman (1994). TBP-DNA interactions in the minor groove discriminate between A:T and T:A base pairs. *Nucleic Acids Res* **22**(10): 1890-6.

Wright, S. I., C. B. Yau, M. Looseley and B. C. Meyers (2004). Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. *Mol Biol Evol* **21**(9): 1719-26.

Wu, F., L. A. Mueller, D. Crouzillat, V. Petiard and S. D. Tanksley (2006). Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics* **174**(3): 1407-20.

Xu, H. and T. R. Hoover (2001). Transcriptional regulation at a distance in bacteria. *Curr Opin Microbiol* **4**(2): 138-44.

Yamamoto, Y. Y., H. Ichida, T. Abe, Y. Suzuki, S. Sugano and J. Obokata (2007). Differentiation of core promoter architecture between plants and mammals revealed by LPSS analysis. *Nucleic Acids Res* **35**(18): 6219-26.

Yamamoto, Y. Y., H. Ichida, M. Matsui, J. Obokata, T. Sakurai, M. Satou, M. Seki, K. Shinozaki and T. Abe (2007). Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* **8**: 67.

Yang, Y. W., K. N. Lai, P. Y. Tai and W. H. Li (1999). Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J Mol Evol* **48**(5): 597-604.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**(8): 1586-91.

Yu, J., S. Hu, J. Wang, G. K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang*, et al.* (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296**(5565): 79-92.

Yu, J., J. Wang, W. Lin, S. Li, H. Li, J. Zhou, P. Ni, W. Dong, S. Hu, C. Zeng*, et al.* (2005). The Genomes of Oryza sativa: a history of duplications. *PLoS Biol* **3**(2): e38.

Zhang, J. and X. He (2005). Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* **22**(4): 1147-55.

Zhao, X. P., Y. Si, R. E. Hanson, C. F. Crane, H. J. Price, D. M. Stelly, J. F. Wendel and A. H. Paterson (1998). Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res* **8**(5): 479-92.

Zimmer, A., D. Lang, S. Richardt, W. Frank, R. Reski and S. A. Rensing (2007). Dating the early evolution of plants: detection and molecular clock analyses of orthologs. *Mol Genet Genomics* **278**(4): 393-402.

Zimmermann, P., L. Hennig and W. Gruissem (2005). Gene-expression analysis and network discovery using Genevestigator. *Trends Plant Sci* **10**(9): 407-9.

Zimmermann, P., M. Hirsch-Hoffmann, L. Hennig and W. Gruissem (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol* **136**(1): 2621-32.

Summary
The main objectives of this thesis were (i) to search for the selection pressure exerted on plant orthologous genes and (ii) to describe the structural and functional features they share in particular in their promoters, in order to (iii) define a novel phylogenetic footprinting approach. Pairs of unique genes, defined by sequence comparisons, have been used because they were considered as having the greater chance to be true orthologues. Plant unique genes form three groups of genes with different structural, functional and evolutionary features. The unique genes that are specific either to *Arabidopsis thaliana* or *Oryza sativa* have features that are different than those of conserved unique genes. On one hand, the species-specific unique genes code preferentially for excreted peptides implied in regulatory functions. On the other hand, the conserved unique genes have characteristics described in genes implied in basal methabolism of the cells and which evolve slowly. Some potential regulatory motifs have been found over-represented only in the promoter of these genes. Lastly, after duplication, these genes generally lose one duplicate, which suggest a strong negative selection against the co-existence of two not diverged copies of the same gene.

Resumen
Los objetivos de la tesis han sido (i) poner de manifiesto la presión de selección que existe sobre los genes ortólogos en plantas y (ii) describir las características estructurales y funcionales que comparten, especialmente a nivel de su promotor, para (iii) definir una nueva aproximación de huella filogenética. Para apoyar el estudio en unas relaciones de ortología lo menos ambiguas posibles, el estudio se ha realizado en aquellos genes 'únicos' en plantas basándonos en una comparación de secuencia. Los genes únicos forman 3 grupos estructuralmente, funcionalmente y evolutivamente diferentes. Tanto los genes únicos específicos de *Arabidopsis thaliana* como los genes únicos específicos de *Oryza sativa* presentan unas características diferentes a la de los genes únicos conservados. Así, por una parte, los genes únicos específicos de una especie son preferencialmente genes que codifican péptidos excretados e implicados en funciones de regulación. Por otra parte, los genes únicos conservados poseen características que están asociadas a genes implicados en funciones de metabolismo de base de las células y que evolucionan lentamente. Ciertos motivos potencialmente reguladores se han encontrado sobre-representados específicamente en el promotor de estos genes. Por otro lado, en general estos genes una vez duplicados vuelven rápidamente a un estado de unicidad, lo que sugiere una fuerte desventaja selectiva a la existencia simultánea de dos copias poco divergentes de estos genes.

Résumé
Les objectifs de la thèse étaient (i) de mettre en évidence les pressions de sélection qui s'exercent sur les gènes orthologues chez les végétaux et (ii) de cerner les caractéristiques structurales et fonctionnelles qu'ils partagent, notamment au niveau de leurs promoteurs, pour (iii) définir une nouvelle démarche d'empreinte phylogénétique. Pour s'appuyer sur des relations d'orthologie les moins ambiguës possibles, l'étude a été centrée sur les gènes 'uniques' définis par comparaison de séquences. Les gènes uniques forment 3 groupes structurellement, fonctionnellement et évolutivement distincts. Les gènes uniques spécifiques soit d'*Arabidopsis thaliana* soit d'*Oryza sativa* présentent des caractéristiques différentes des gènes uniques conservés, y compris au niveau de leur promoteur. D'une part, les gènes uniques spécifiques à une espèce sont préférentiellement des gènes codant pour des peptides excrétés et impliqués dans des fonctions de régulation. D'autre part, les gènes uniques conservés ont des caractéristiques qui les associent aux gènes impliqués dans les fonctions du métabolisme de base des cellules et qui évoluent lentement. Certains motifs potentiellement regulateurs ont été trouvés spécifiquement sur-representés dans leur promoteur. Par ailleurs, ces gènes, quand ils sont dupliqués, retournent rapidement à l'état unique, suggérant un fort désavantage sélectif à l'existence simultanée de deux copies peu divergentes de ces gènes.

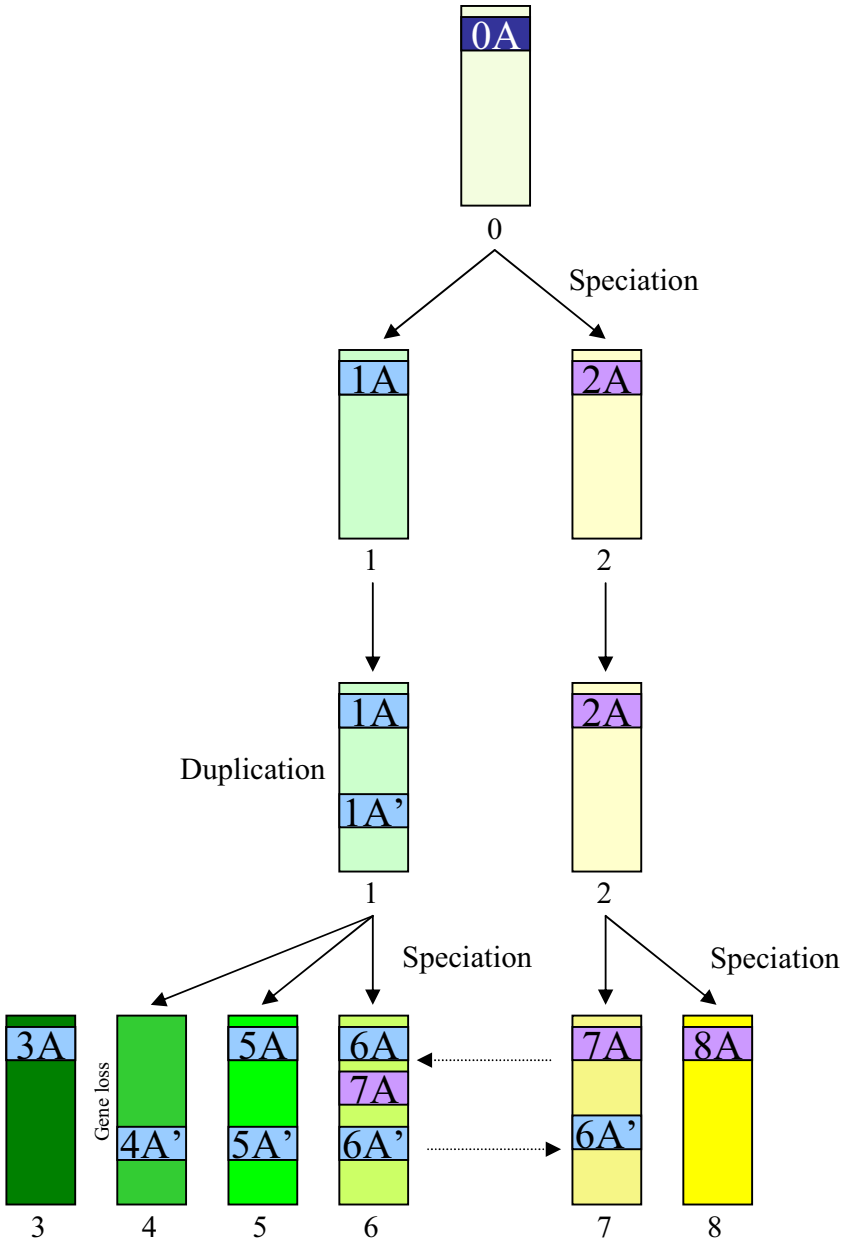

Figure 1 - Orthologs, paralogs and other relationships
Schematic illustration representing the different relationships that can be defined between the genes. The same gene A is represented by a coloured small box contained within different species genomes. Each species are also coloured differently and numbered to make them more distinguishable while duplicated genes are denoted by an apostrophe. This illustration includes examples of orthology (1A-2A), paralogy (1A-1A'), in-paralogy (1A and 1A'-2A), out-paralogy (5A and 5A'-3A), pseudo-orthology (3A-4A') and xenology (6A' in species 7A).

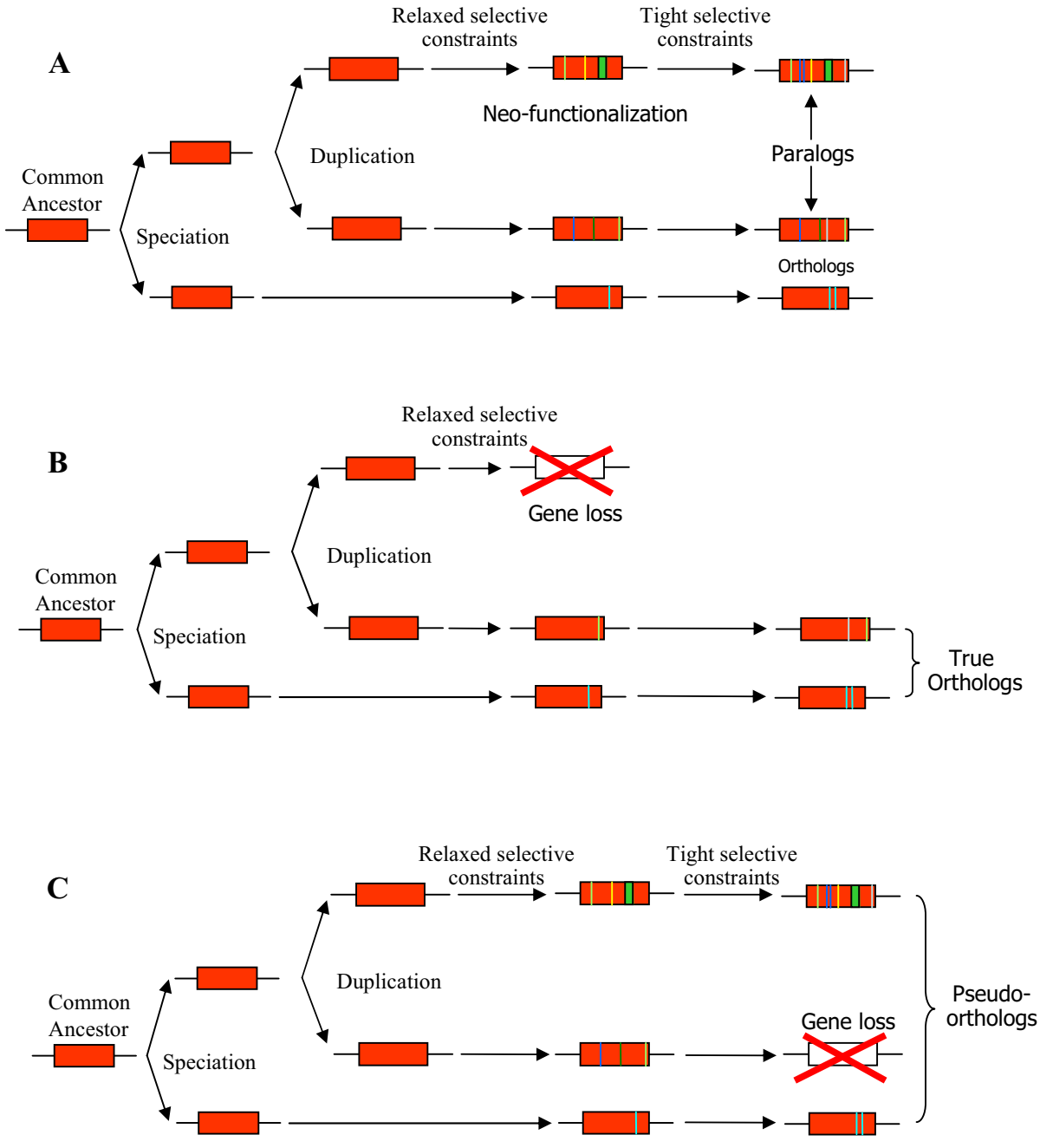

Figure 2 – True orthologs and pseudo-orthologs
Schematic illustration representing the relationship evolution of an ancestral gene in two different species after a duplication event: (A) In absence of gene loss, one of the duplicates will probably be free to accumulate mutations and acquire a new function. Within this situation the duplicated genes would be defined as paralogs while the gene accumulating less mutations will have an orthology relationship with the gene in other species evolved from common ancestor. (B) If soon after duplication, one of the duplicated genes is lost, the remaining gene would not have had time to accumulate mutations and will have a true orthology relationship with the gene in the other species. (C) However, if gene loss occurs a long time after duplication event in such a way that one of the genes has had time to acquire new function and it is precisely this gene which is retained, it will be defined as ortholog of the gene in the other species instead of pseudo-orthologs.

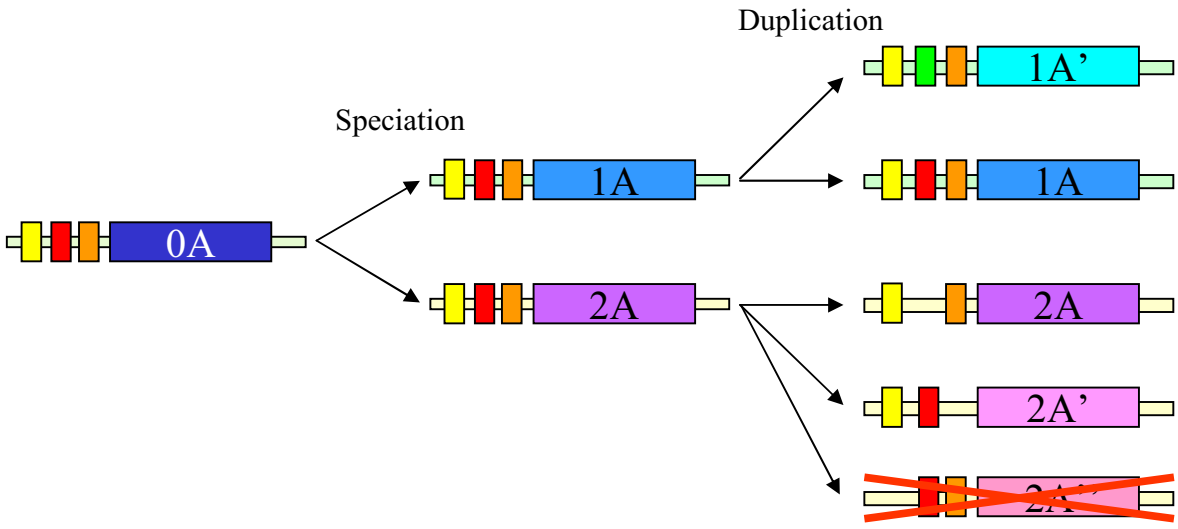

Figure 3 – Sub-, Neo- and No-functionalization
Schematic illustration representing the different consequences that can have the loss or gain of different regulatory factors and/or protein subunits. The same gene A is represented by a coloured box preceded by different regulatory motifs with duplicated genes denoted by an apostrophe. This illustration includes examples of sub-functionalization (2A-2A'), neo-functionalization (1A') and no-functionalization (2A'').

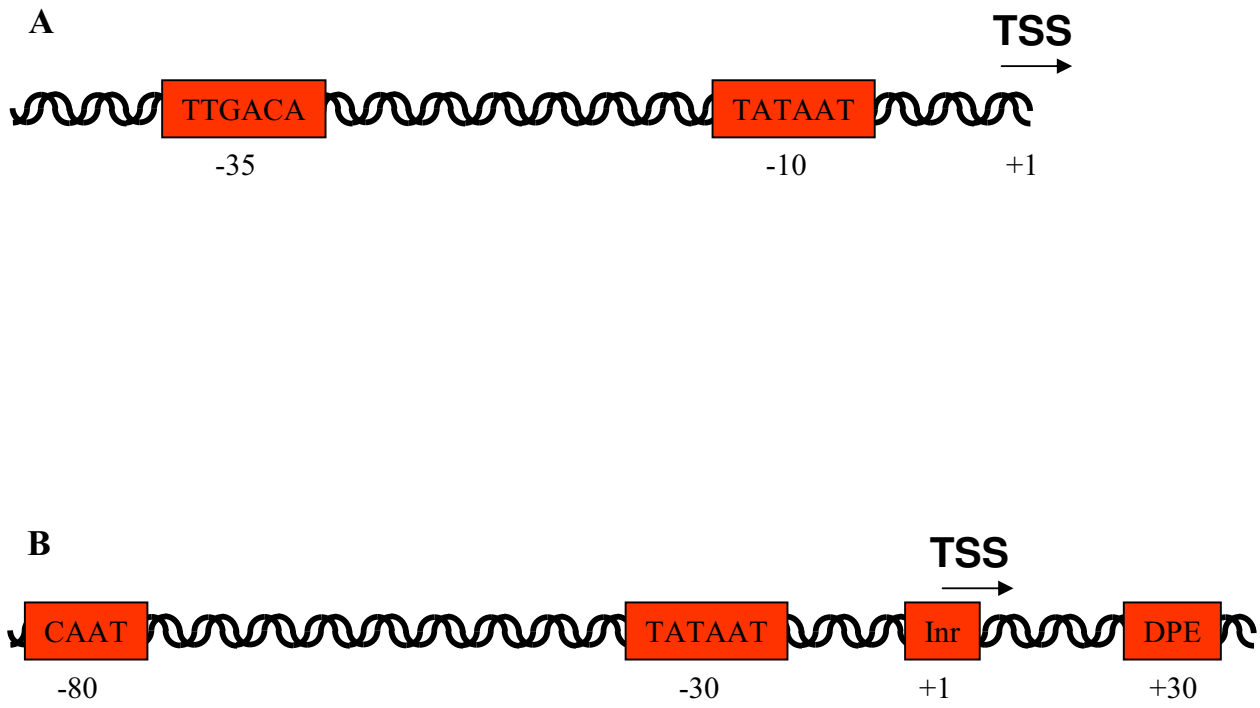

Figure 4 – Prokaryotic and eukaryotic promoters
(A) Schema of a classic prokaryotic promoter including the Pribnow box (TATAAT) and TTGACA motif at their consensus position with respect to the TSS. (B) Schema of a classic eukaryotic promoter including TATA-box, CAAT-box, Inr and DPE at their consensus position with respect to the TSS. Despite these motifs are the classically described in the literature to illustrate the organization of promoters, they can present many variability including new TFBS, absence of the classical TFBS and differences in the position with respect to the TSS.

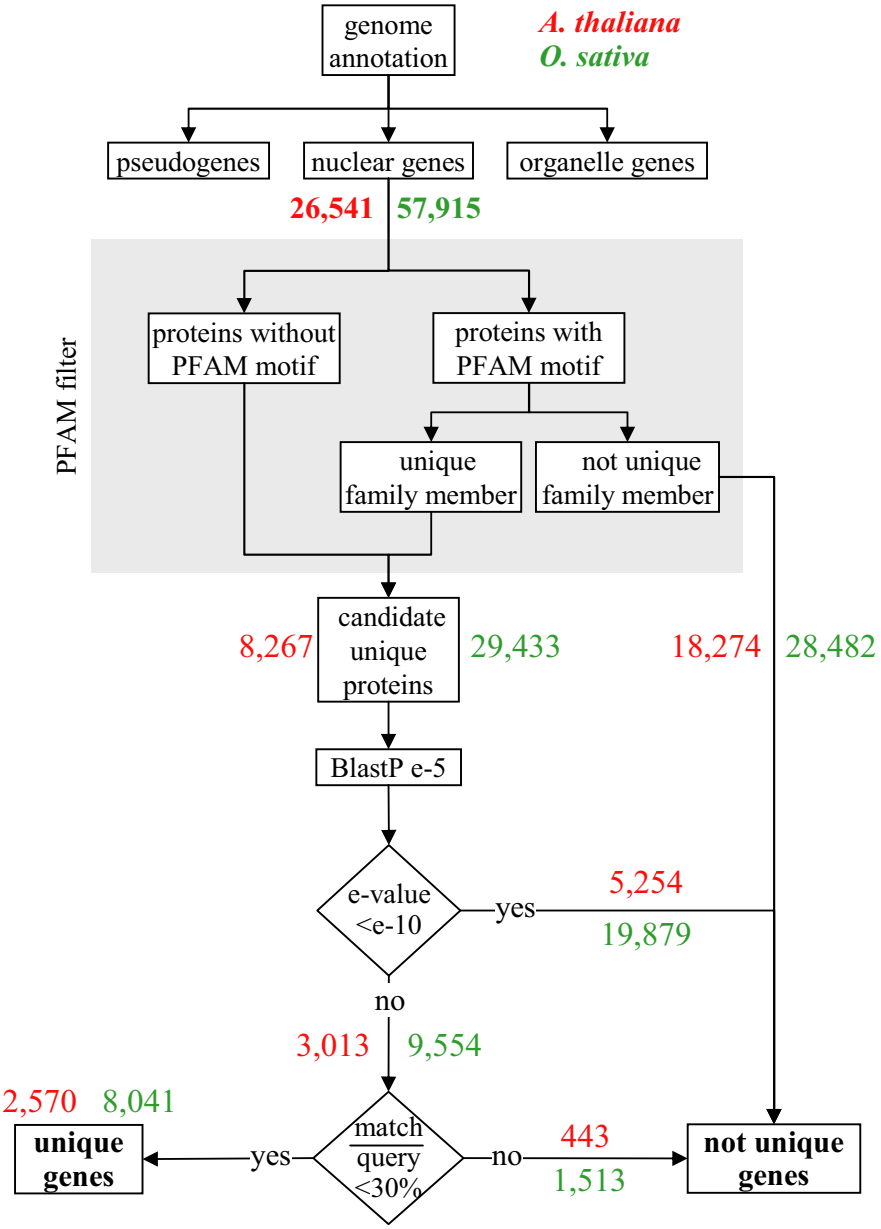

Figure 5 – Characterization of unique genes in *A. thaliana* and *O. sativa*
Schematic diagram describing the different filters applied to obtain the list of unique genes in each species. Only the proteins encoded by the nuclear genes were used. PFAM filter removed members of known families and BLASTp filters eliminated other genes with at least one homolog in the same genome. Results from *A. thaliana* genome are labelled in red while *O. sativa* results are in green.

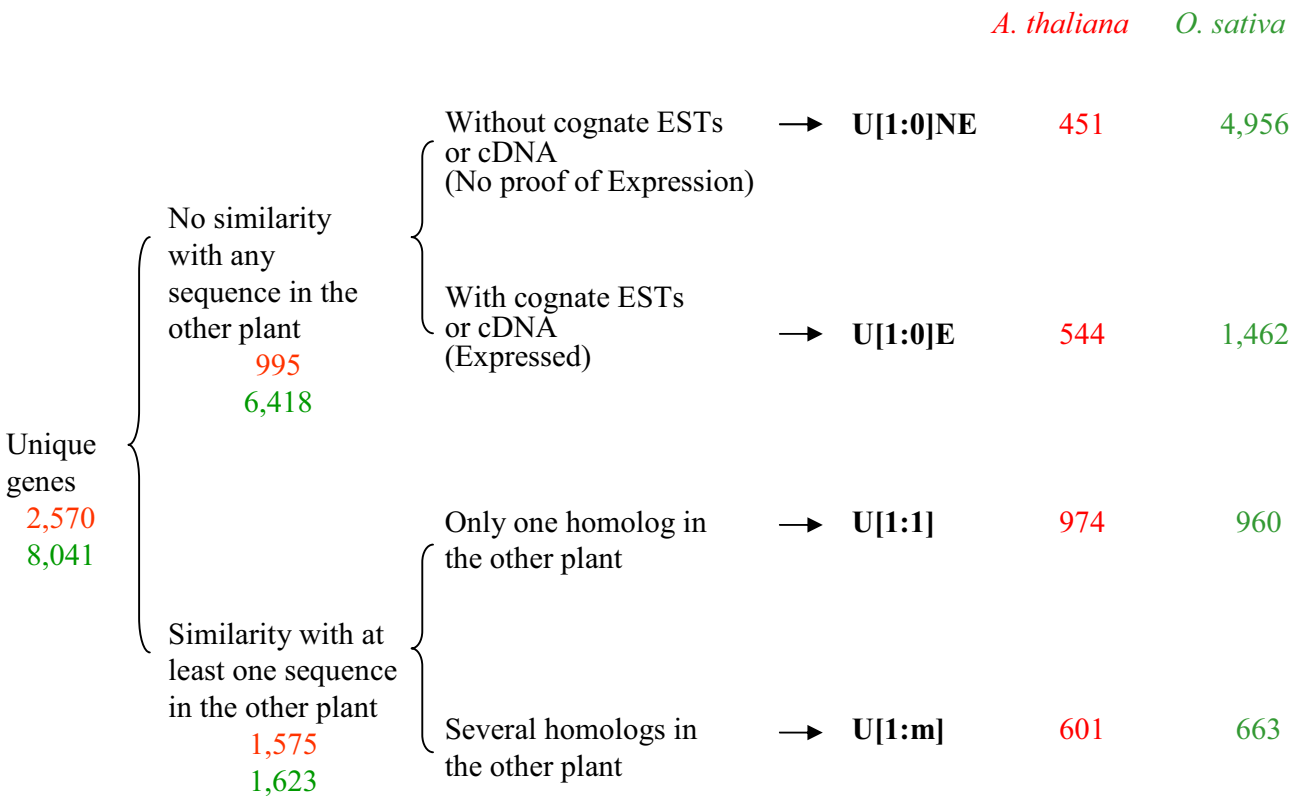

Figure 6 – Unique genes classification

Based on BLASTp sequence comparison, *A. thaliana* and *O. sativa* unique genes were classified according to the number of homologs in the other species. We named U[1:0] the unique proteins in one species with no homolog in the other one, U[1:1] the unique proteins with only one homolog and U[1:m] the unique proteins with more than one homolog. First, a BLASTp between unique protein in each species and the whole proteome of the other species was used to define U[1:0], U[1:1] and U[1:m] gene groups. Proofs of transcription (presence of cognate ESTs and/or cDNA) were used for further classification of U[1:0] genes in U[1:0]E (for Expressed) and U[1:0]NE (for No proof of Expression) genes. Red numbers are relative to *A. thaliana* while green ones are relative to *O. sativa*.

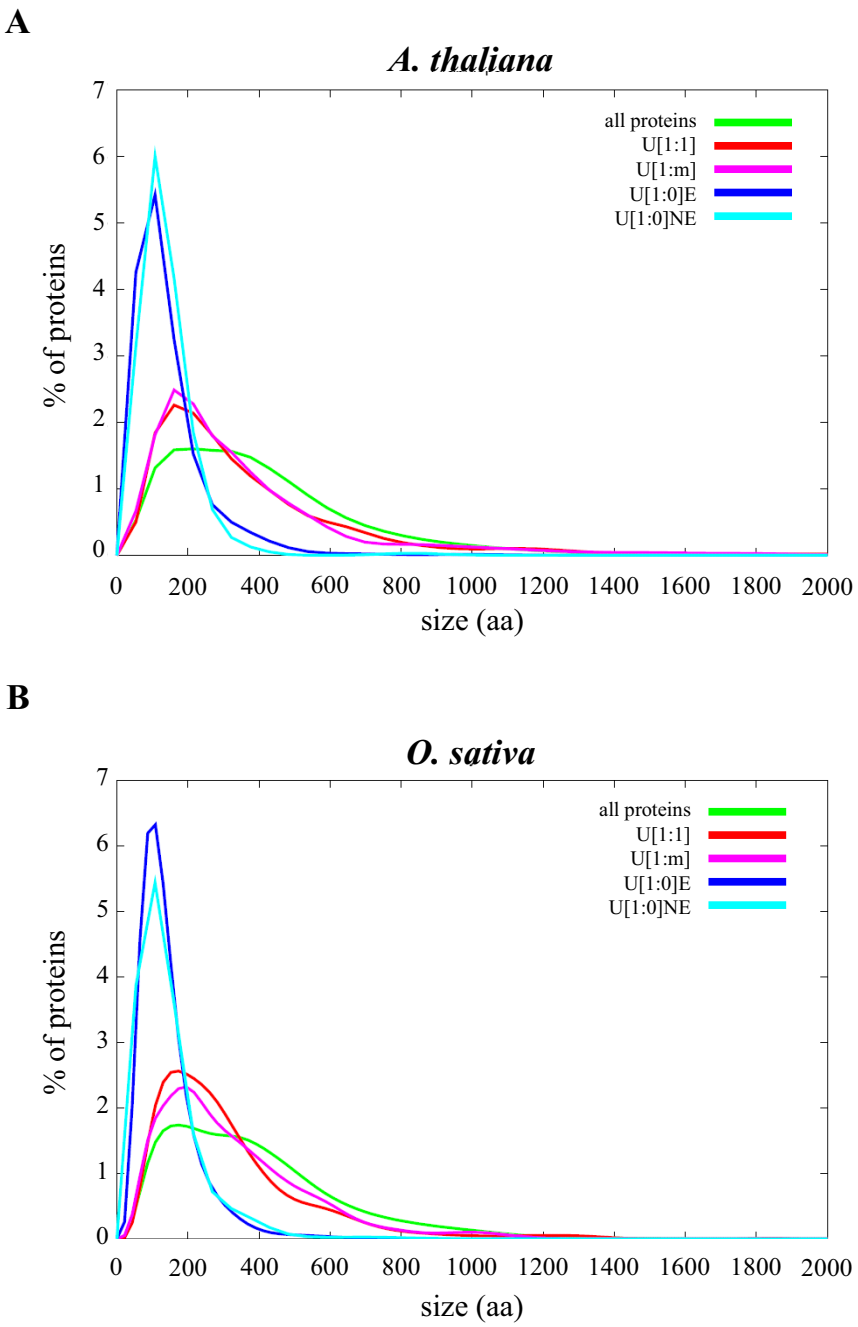

Figure 7 – Size distributions of proteins encoded by unique genes
The size distributions of different groups of proteins encoded by unique genes are compared in *Arabidopsis thaliana* (A) and *Oryza sativa* (B). The reference 'all proteins' corresponds to every proteins encoded by the nuclear genes.

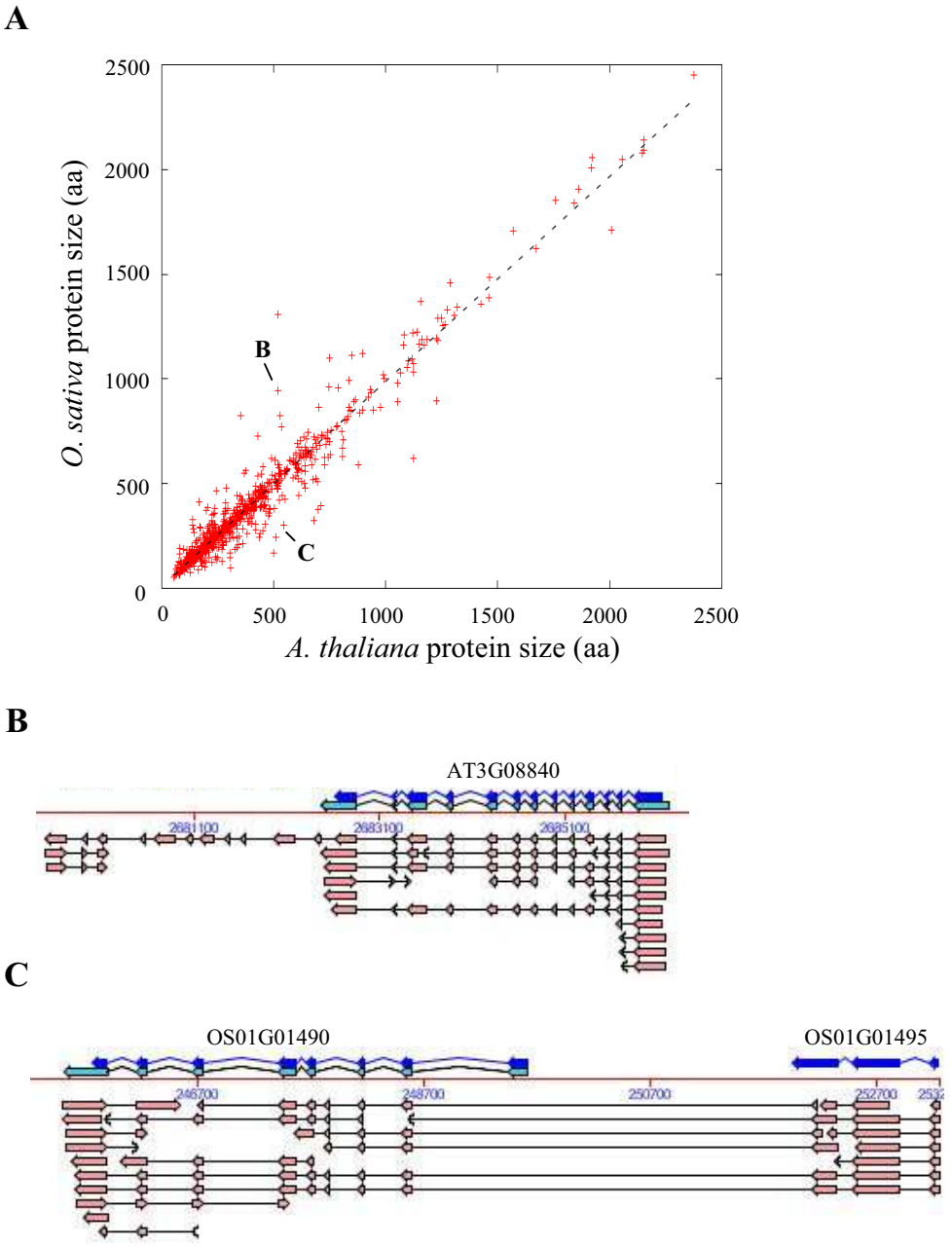

Figure 8 – Comparison of protein lengths in U[1:1] pairs

Each point represents protein lengths (in aa) of one U[1:1] pair of proteins (A). The linear correlation between U[1:1] protein sizes is represented by a dotted line (r2=0.94). Hand-checking of the largest differences showed that they are mainly due to erroneous predicted gene models with either an artificial exon gain/loss as in AT3G08840 (B) or a splitting/fusion process as in OS01G01490-OS01G01495 (C). Arrows and lines represent exons and introns while dark blue, light blue and pink colours represent predicted CDS, predicted mRNA and cognate transcripts (ESTs/cDNA), respectively. (B) and (C) are snapshots from FLAGdb^{++} (Samson *et al.*, 2004).

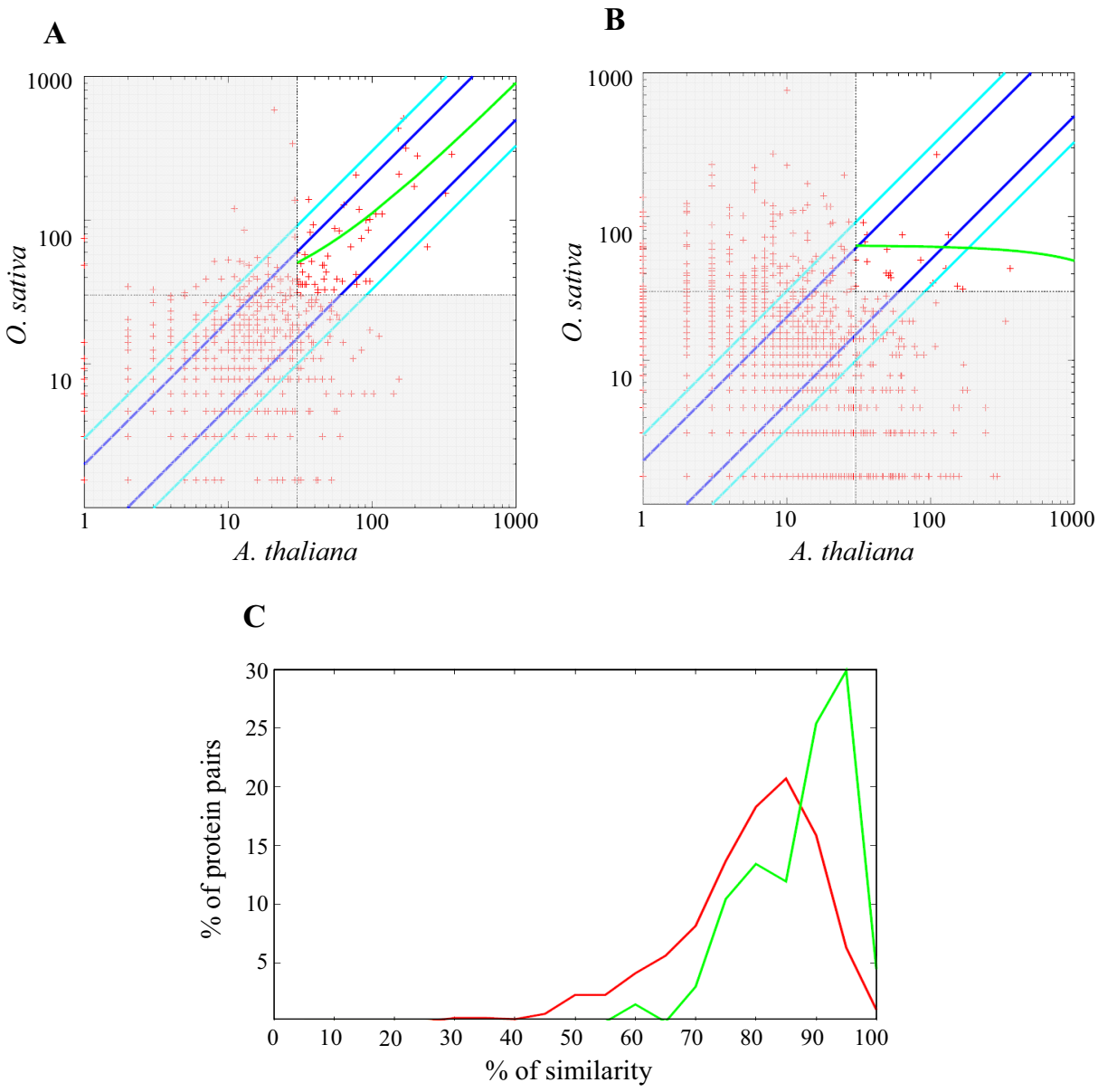

Figure 9 – Expression levels correlated between genes of U[1:1] pairs
Expression level correlation based on the number of transcripts (ESTs/cDNA) associated to U[1:1] gene pairs (A) and randomized nuclear gene pairs (B). Values were first normalized to take into account the size of the transcript resources in each species, the number of genes with a transcript and the total number of genes on each species, and then transformed by base 10 logarithm. We used only the gene pairs with a size difference between proteins equal to or smaller than 20 aa (526 U[1:1] and 8,390 randomized pairs). The green line represents the linear correlation for pairs of genes with at least 30 cognate transcripts (white area). U[1:1] genes pairs: r2=0.51 and Kendall's test P-value=1e-6; Random pairs sample: r2=0.03 and Kendall's test P-value=0.26. Diagonal lines delimit an expression similarity of 33 % (light blue) and 50 % (dark blue). (C) Percentage of similarity was recovered from ClustalW alignments of U[1:1] protein pairs encoded by highly (green, more than 30 cognate transcripts) and lowly (red, less than 30 cognate transcripts) transcribed genes.

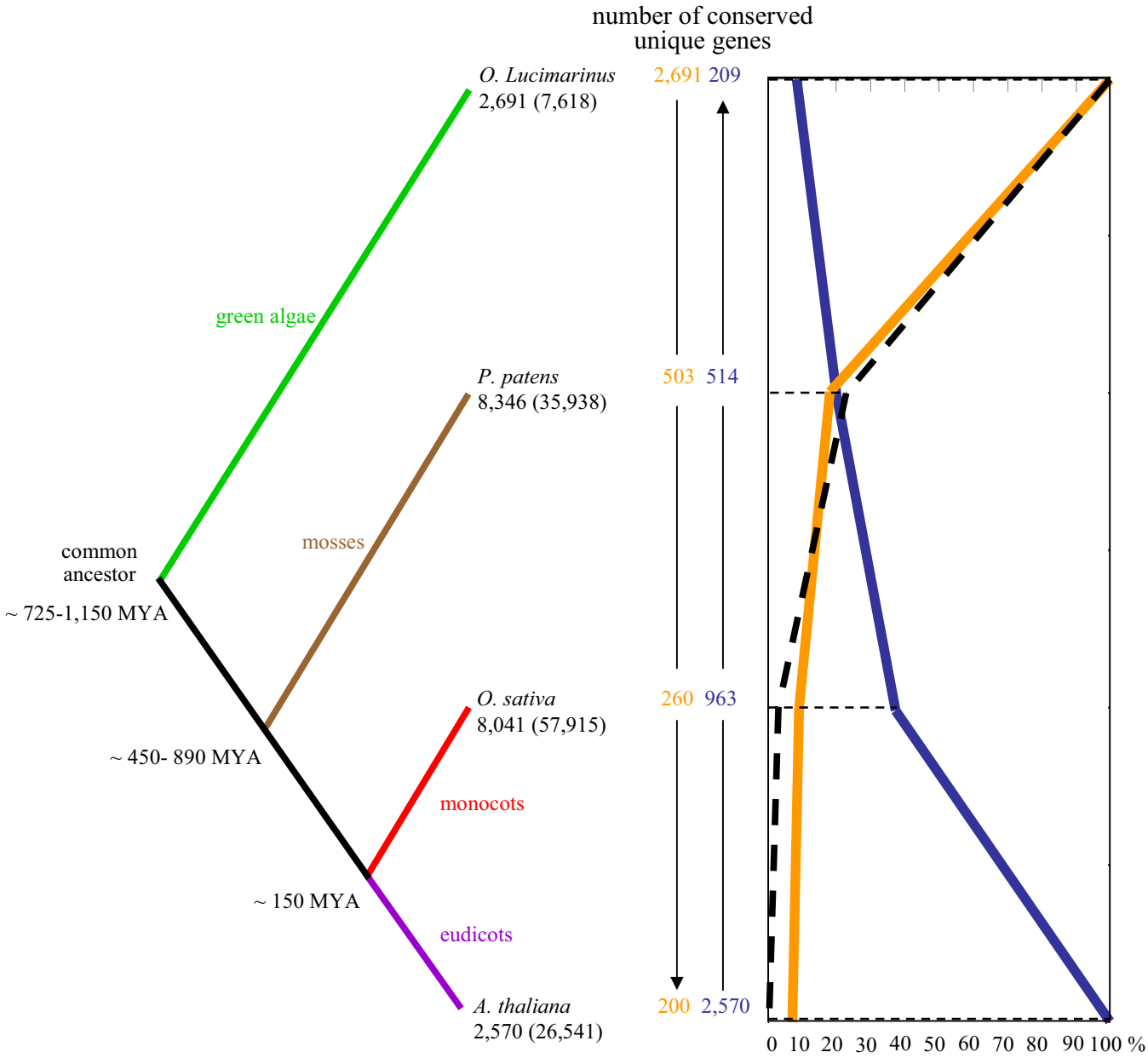

Figure 10 – Unique gene conservation in the plant kingdom

Study of unique gene conservation through evolution of *Arabidopsis thaliana* (Brassicales), *Oryza sativa* (Poales), *Physcomitrella patens* (Funariaceae) and *Ostreococcus lucimarinus* (Prasinophyceae). Unique genes of each species were characterized (number below species name, total nuclear genes between brackets) and orthology relationships between couples of species were established using the previously described protocol. Phylogenetic conservation of unique genes was analysed from *O. lucimarinus* (orange line) and *A. thaliana* (blue line) discarding not conserved unique genes on each node (evolution distance showed in millions of years) (Wolfe *et al.*, 1989; Chaw *et al.*, 2004; Hedges *et al.*, 2004; Zimmer *et al.*, 2007; Rensing *et al.*, 2008). Remaining genes in each case were compared to eliminate inconsistencies and obtain a final list of 192 unique genes conserved as unique in the four species: U[1:1:1:1] genes. These 192 conserved unique genes are far more than the 8.38 U[1:1:1:1] genes expected by random conservation (black dashed line).

Figure 11 – CREDO example

Example of a typical CREDO output. Each species promoter is separately represented alongside with the conserved motifs detected by each method of the package. We have signalled in green our positive test alongside with a conserved motif in both species, coloured in pink, which has been detected by different methods and can be therefore considered as significant.

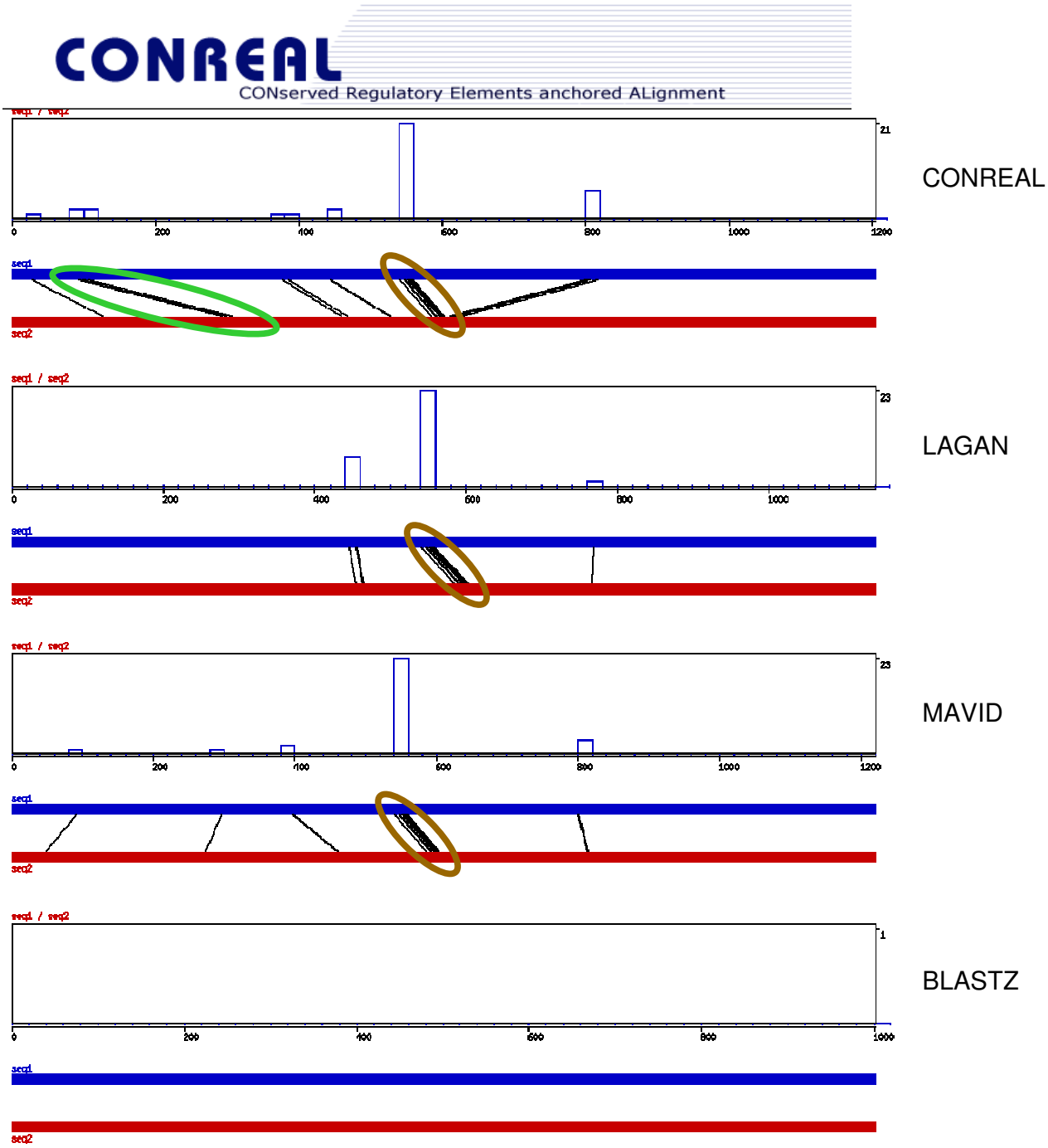

Figure 12 – CONREAL example

Example of a typical CONREAL output. Each method result is independently displayed for each species promoter and a line to indicate the common motif found. In this case we signalled in green our positive test, which could only be recovered by CONREAL method itself, and in brown another conserved motif found by three of the methods that could be considered therefore as significant.

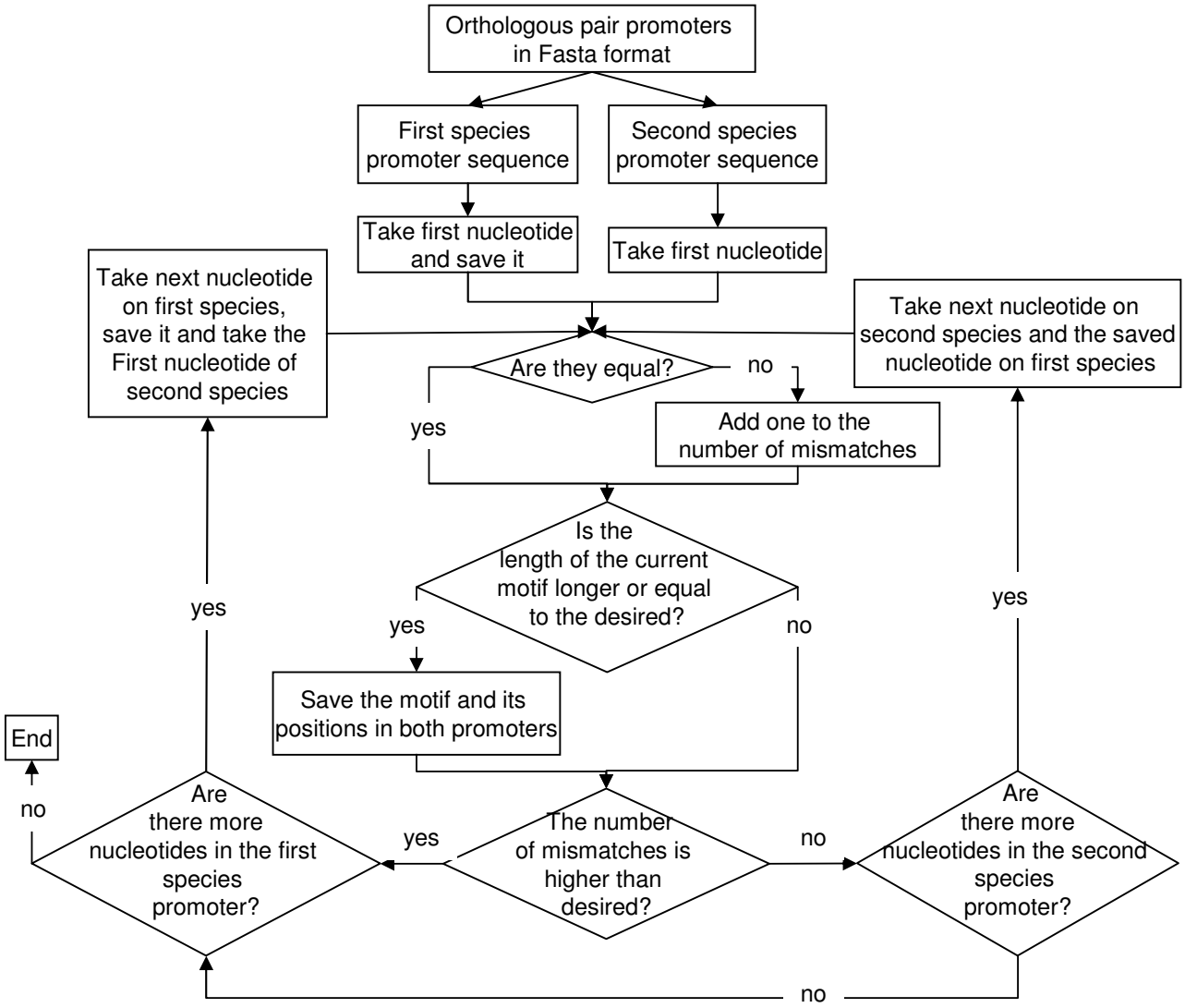

Figure 13 – DECOMO protocol
Schema of the DECOMO protocol employed to detect all the conserved motifs within two sequences based only on sequence comparison and maximum number of mismatches.

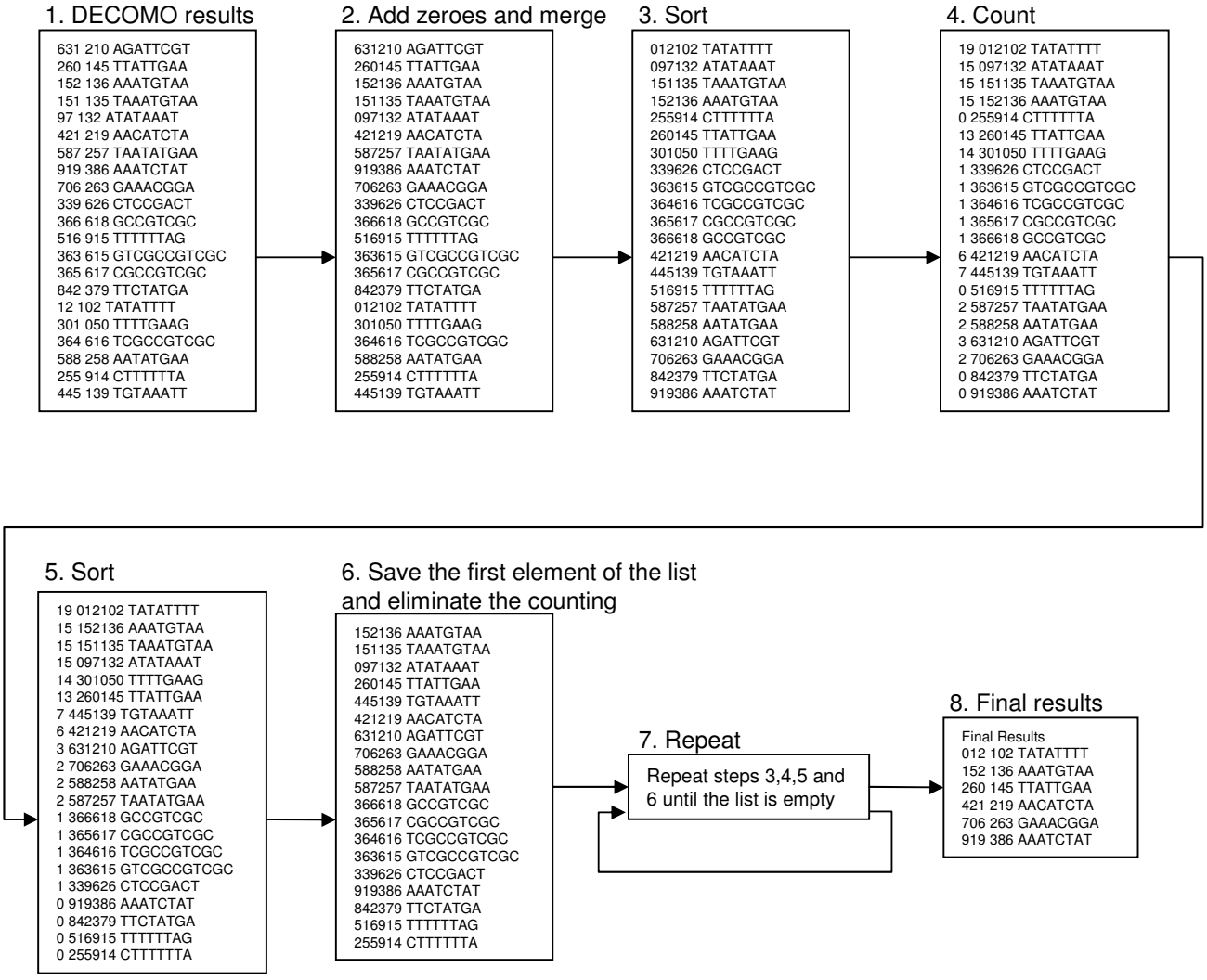

Figure 14 – The 'Maximum synteny' filter algorithm

Output example of the protocol used to find the maximum number of ordered motifs between two promoters. The showed example corresponds to the actual output for the analysis of AT5G44710-OS05G05560 at each step of the protocol. The obtained results show the maximum number of ordered motifs and one of the possible combinations to obtain it.

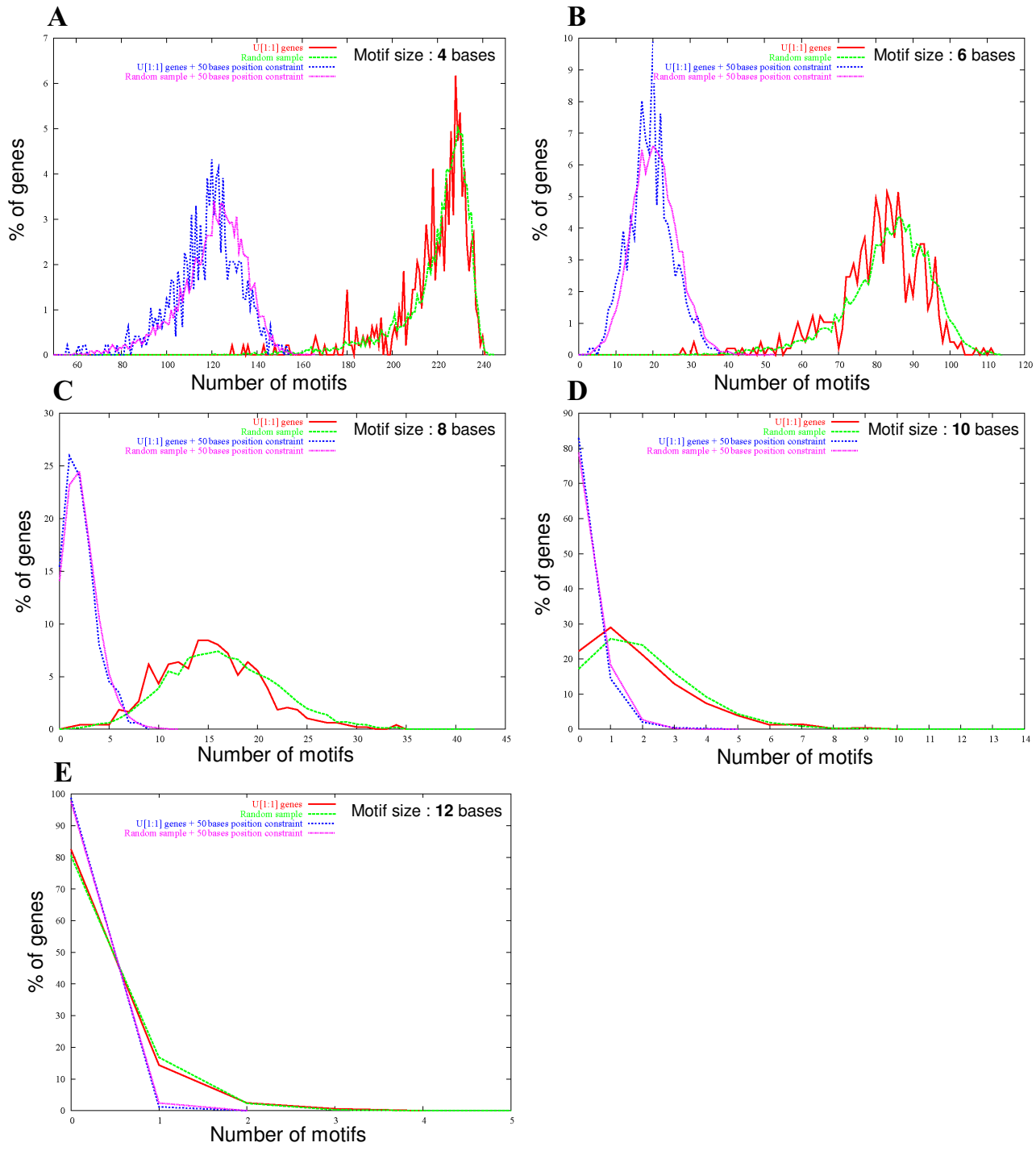

Figure 15 – Number of shared motifs

Number of shared motifs within the promoters of U[1:1] and randomly paired nuclear genes with a confident TSS. Different motifs sizes with no mismatches were tested including motifs of size 4 (A), 6 (B), 8 (C), 10 (D) and 12 (E) nucleotides. The obtained results were filtered with a position constraint filter to select only conserved motifs in similar positions in both species. Motif position constraint filter was tested for a maximum difference of 50, 100 and 200 nucleotides. The graphs show only the results with no and 50 nucleotides position constraint (the 100 and 200 nucleotides position constraints show intermediary results).

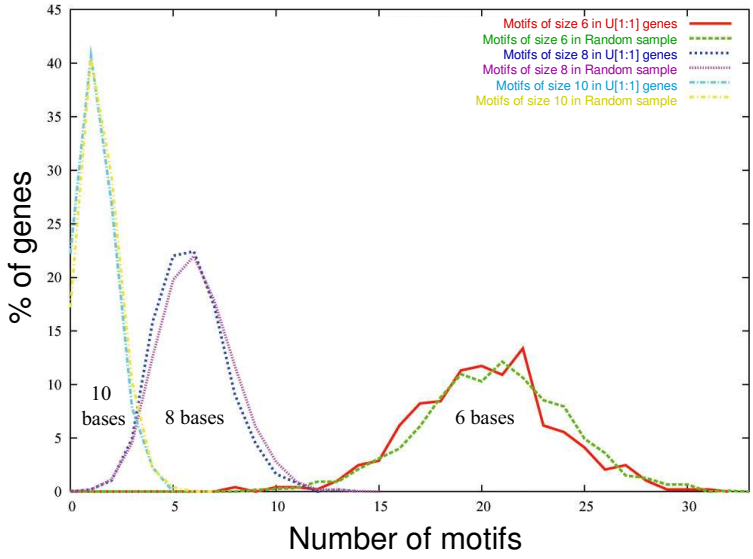

Figure 16 – 'Maximum synteny'
Maximum number of shared ordered motifs within the promoters of U[1:1] genes. Different motif sizes with no mismatch were tested including motifs of size 6, 8 and 10 nucleotides. Motif conservation was compared between U[1:1] and randomly paired nuclear genes for which a confident TSS has been defined.

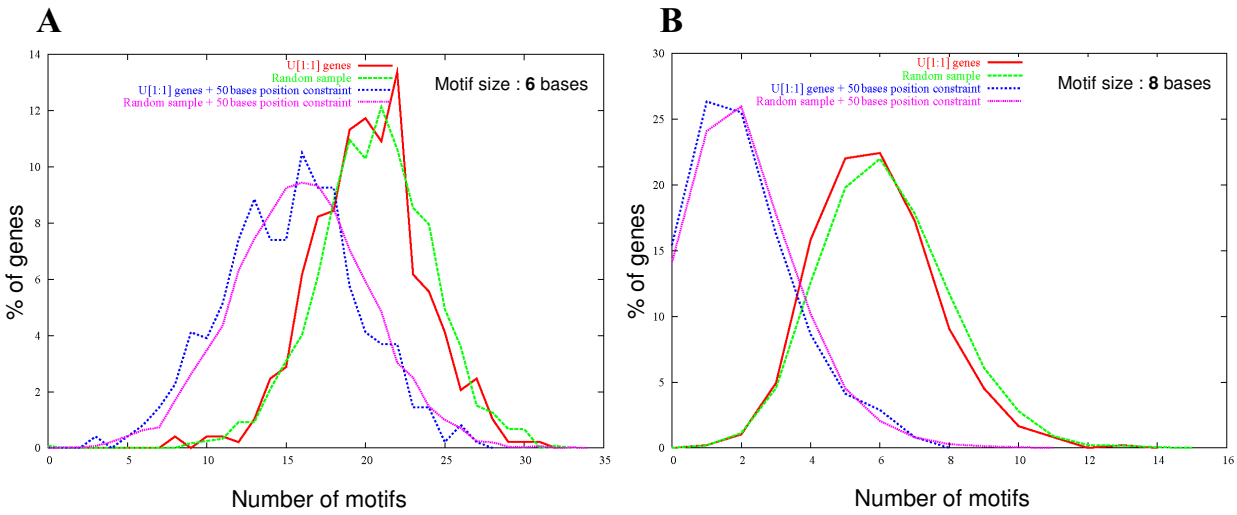

Figure 17 – 'Maximum synteny' combined with position constraint
Maximum number of shared ordered motifs within the promoters of U[1:1] genes. Motifs of sizes 6 (A) and 8 nucleotides (B) with no mismatch were tested. Motif conservation was compared between U[1:1] and randomly paired nuclear genes for which a confident TSS has been defined. The obtained results were filtered with a position constraint filter to select only conserved motifs in similar positions in both species. Motif position divergence filter was tested for a maximum difference of 50, 100 and 200 nucleotides. The graphs show only the results with no and 50 nucleotides position constraint (the 100 and 200 nucleotides position constraints show intermediary results).

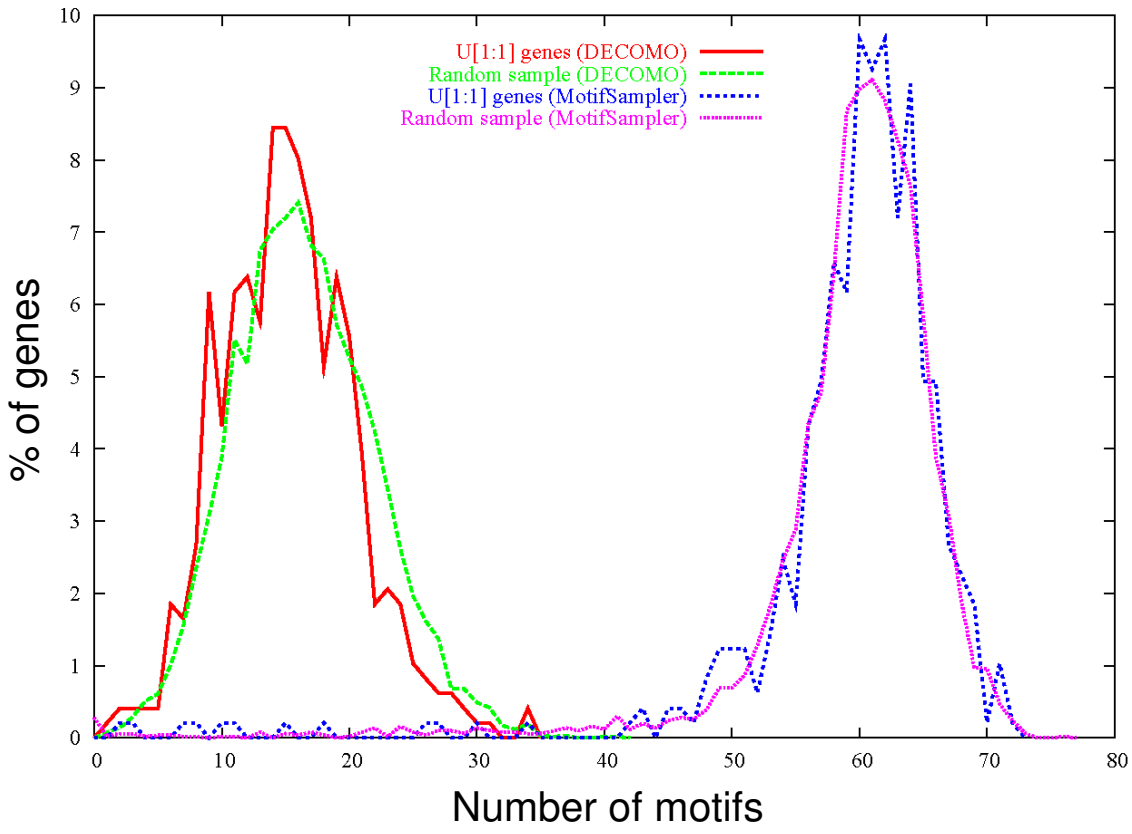

Figure 18 – MotifSampler results compared with DECOMO results
MotifsSampler results for shared motifs of 8 nucleotides with divergence within U[1:1] and randomly paired genes for which a confident TSS has been defined compared with the results obtained for the same samples with DECOMO protocol.

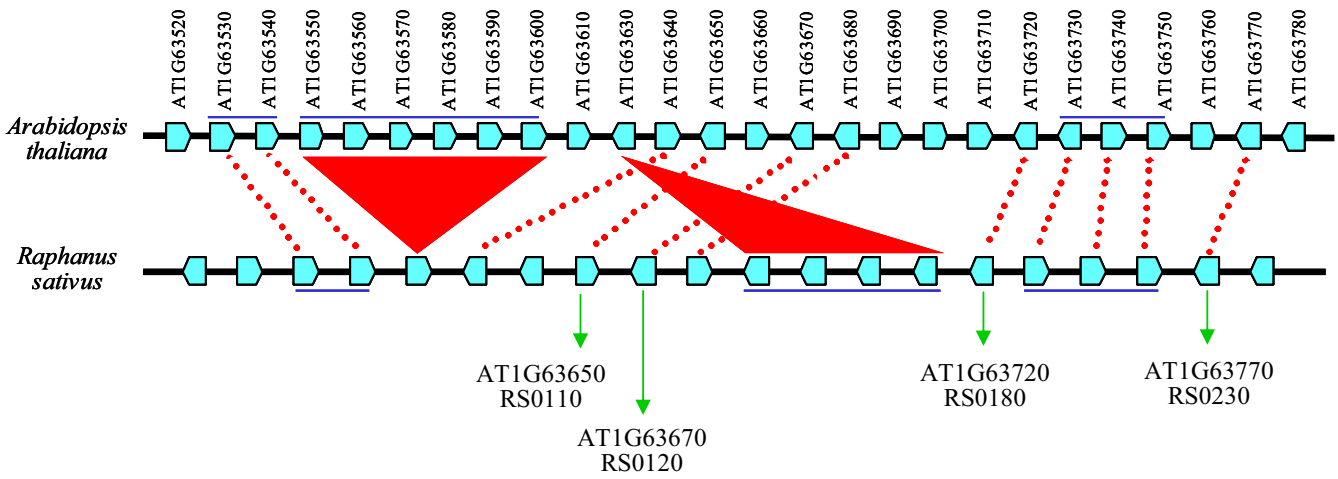

Figure 19 – Synteny between *Arabidopsis thaliana* and *Raphanus sativus*
Schema illustrating the detailed conservation of synteny between the *Arabidopsis* region found by FiToCoGene and a radish BAC sequence. The genes duplicated in tandem are underlined in blue. The orthologous relationships are represented in red. The 4 pairs of genes selected for the positive control are indicated by the green arrows.

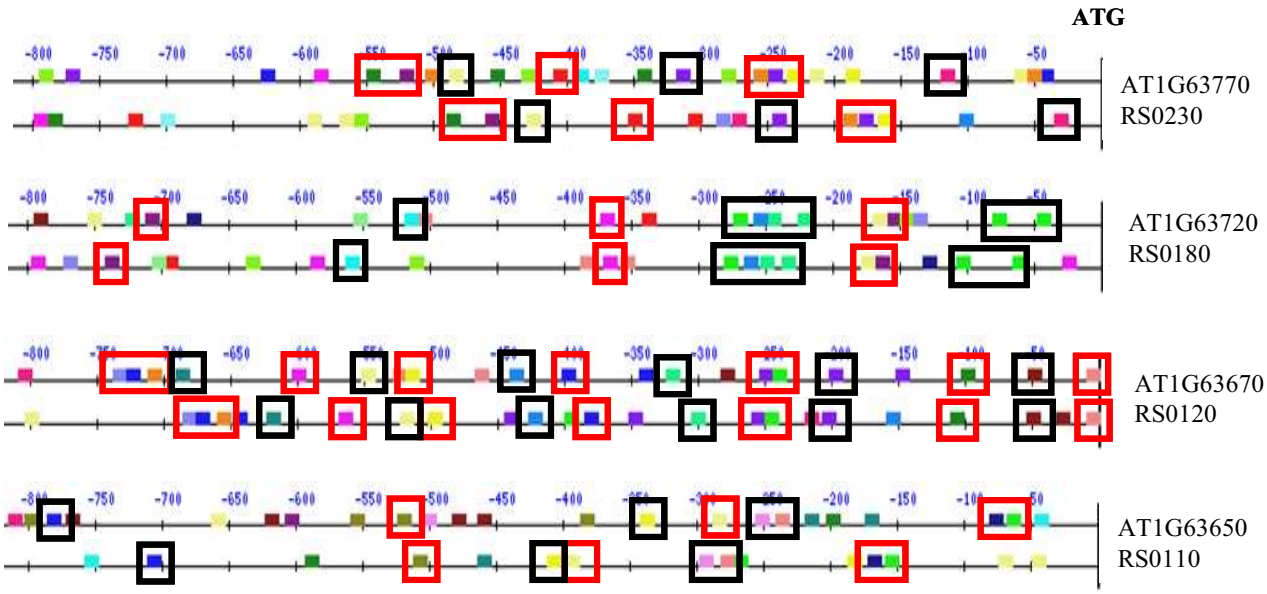

Figure 20 – Conserved motifs in four *Arabidopsis thaliana-Raphanus sativus* gene pairs
Conserved motifs found in the four *Arabidopsis thaliana-Raphanus sativus* gene pairs used as positive test. Each motifs obtained from a single run of MotifSampler are colored with different colours for illustration purposes and ordered conserved motifs were highlighted by hand.

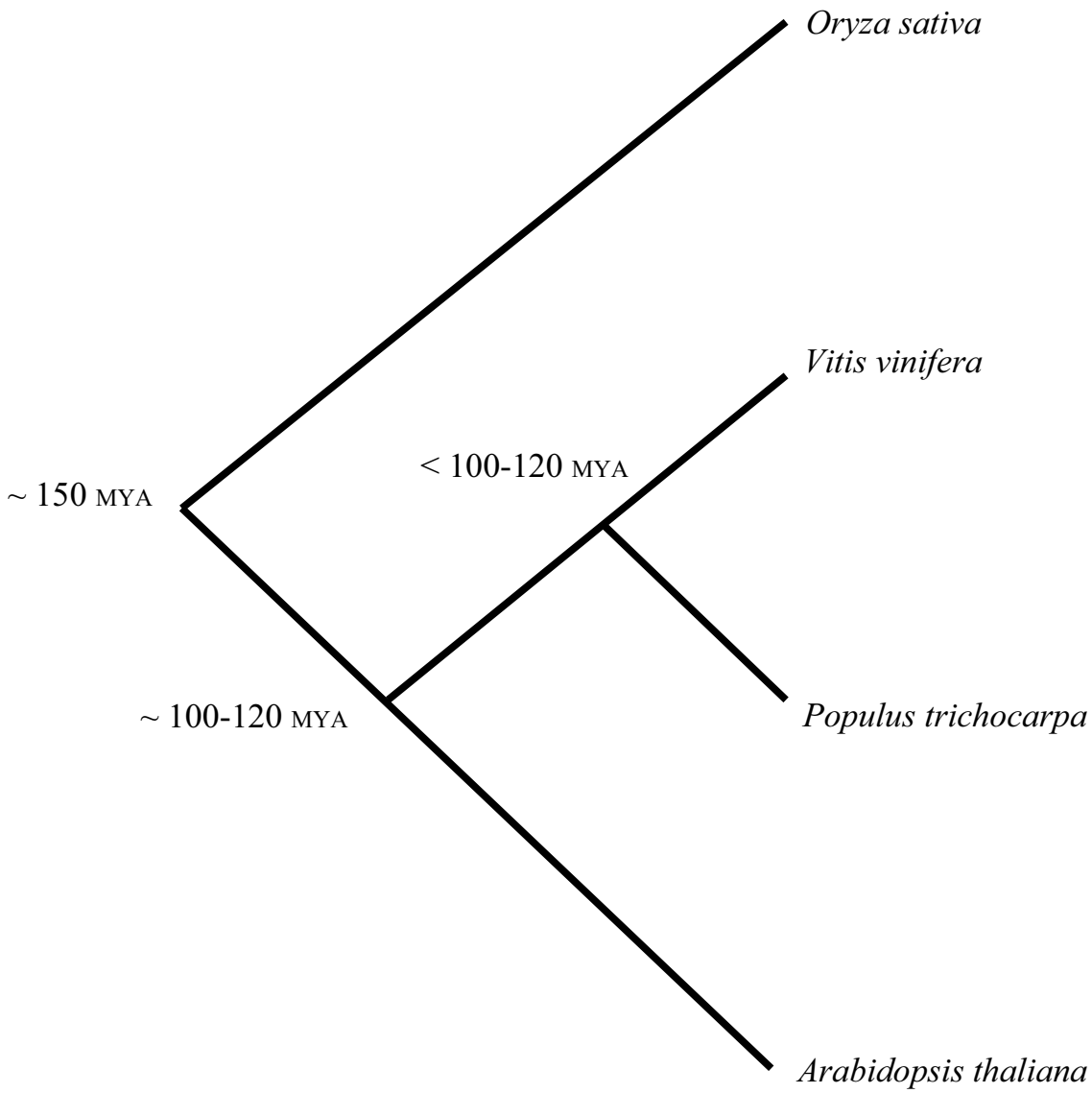

Figure 21 – Schematic cladogram representing the origin of the studied species
The estimated age from radiation is showed on each node according to data in the literature (Wolfe *et al.*, 1989; Chaw *et al.*, 2004; Tuskan *et al.*, 2006, Jaillon *et al.*, 2007).

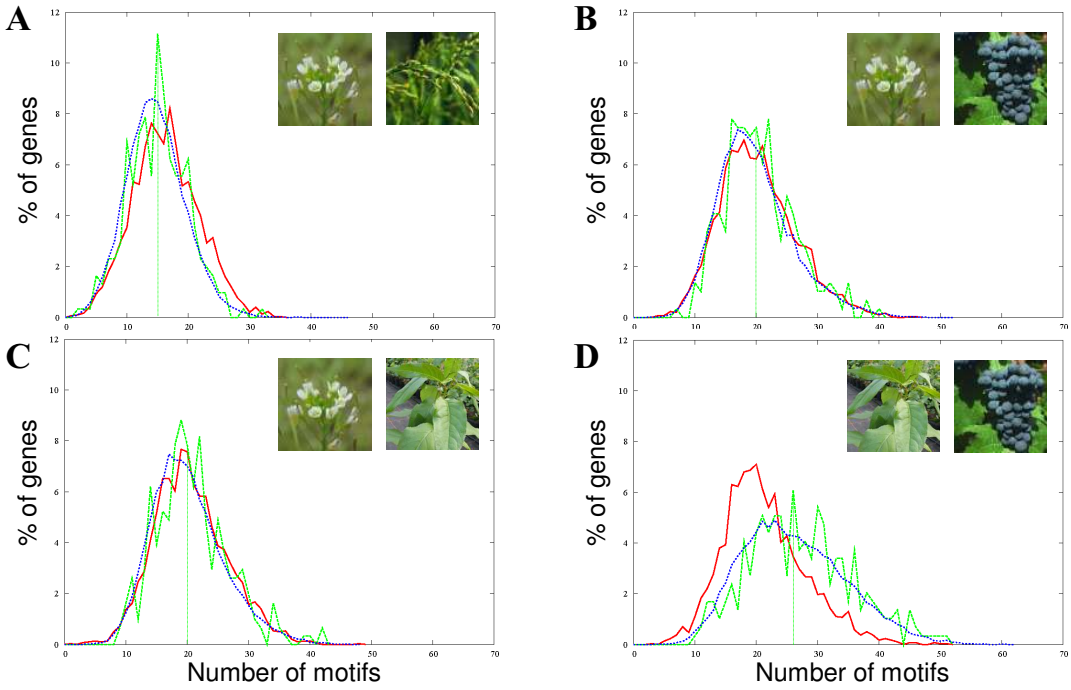

Figure 22 – Number of shared motifs in other species

Number of shared motifs in the promoters of U[1:1] genes (green), randomly paired U[1:1] genes (blue) and randomly paired nuclear genes (red) for which a confident TSS has been defined. The compared species included *A.thaliana-O.sativa* (A), *A.thaliana-V.vinifera* (B), *A.thaliana-P.trichocarpa* (C) and *V.vinifera-P.trichocarpa* (D). Medians are indicated by vertical green line for U[1:1] promoters.

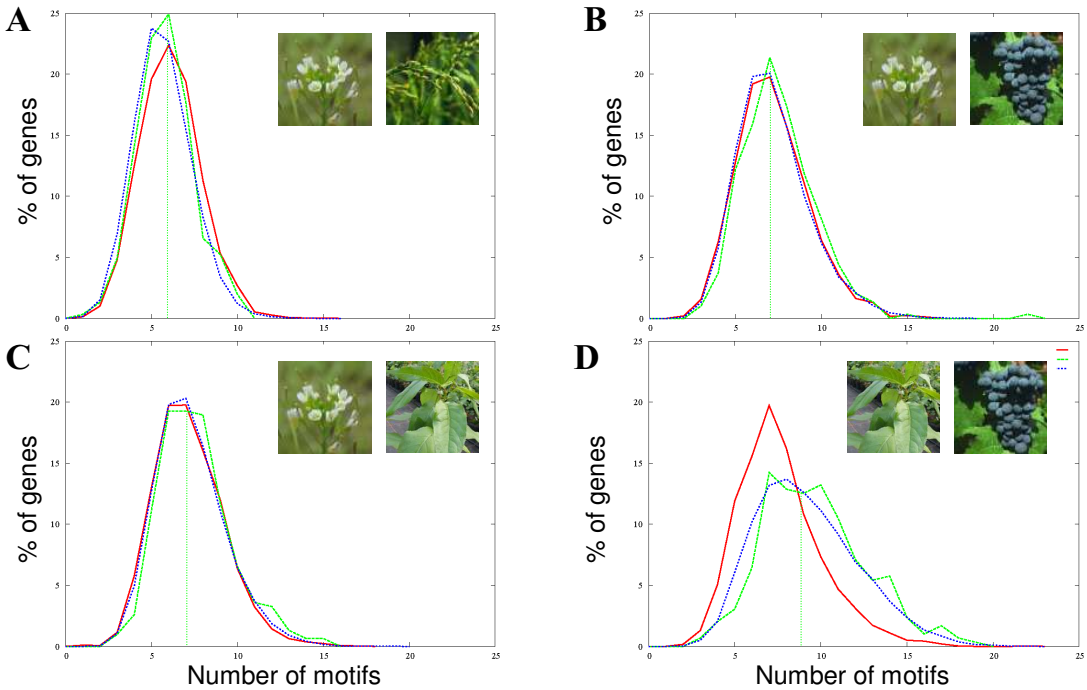

Figure 23 – 'Maximum synteny' in other species

Maximum number of shared ordered motifs of size 8 compared between the promoters of U[1:1] genes (green), randomly paired U[1:1] genes (blue) and randomly paired nuclear genes (red) for which a confident TSS has been defined. The compared species included *A.thaliana-O.sativa* (A), *A.thaliana-V.vinifera* (B), *A.thaliana-P.trichocarpa* (C) and *V.vinifera-P.trichocarpa* (D). Medians are indicated by vertical green line for U[1:1] promoters.

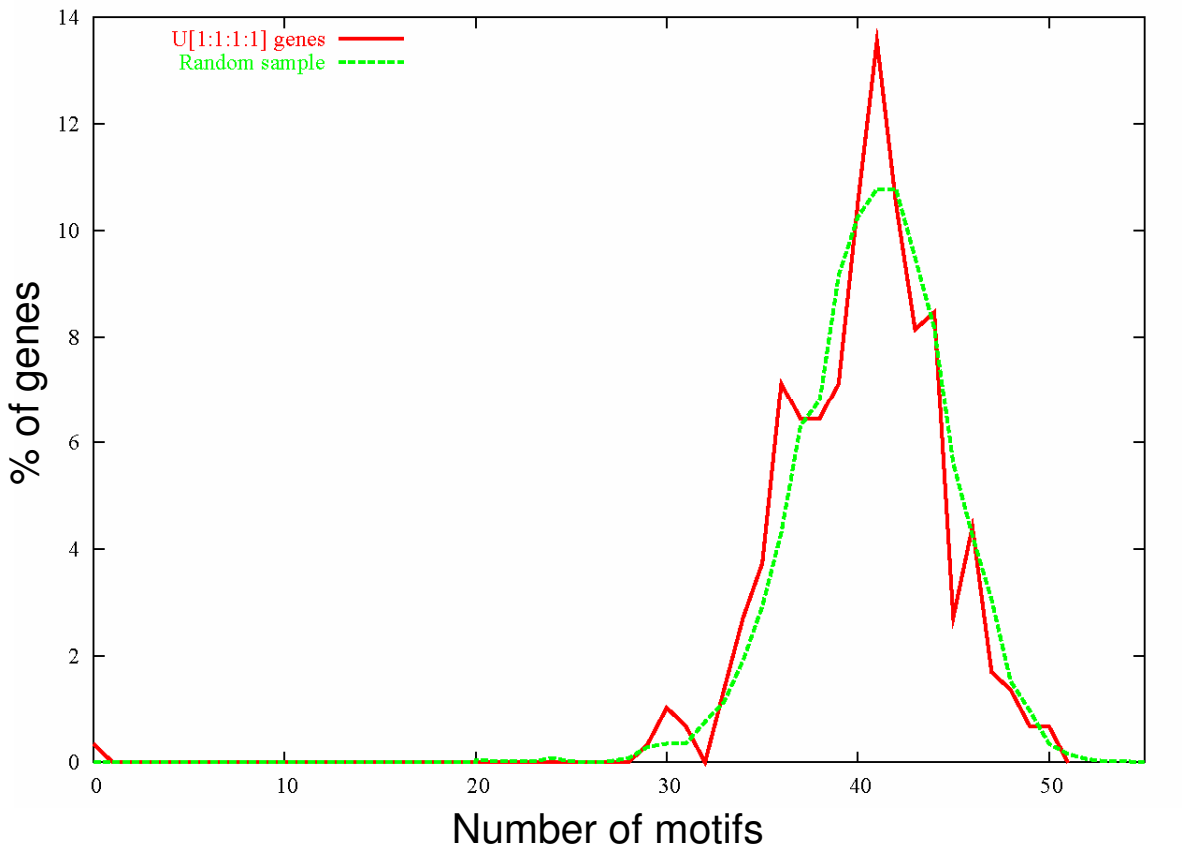

Figure 24 – MotifSampler results for U[1:1:1:1] genes
MotifsSampler results for shared motifs of 8 nucleotides with divergence within U[1:1:1:1] and randomly clustered genes for which a confident TSS has been found. Only motifs detected in the four species promoters have been plotted.

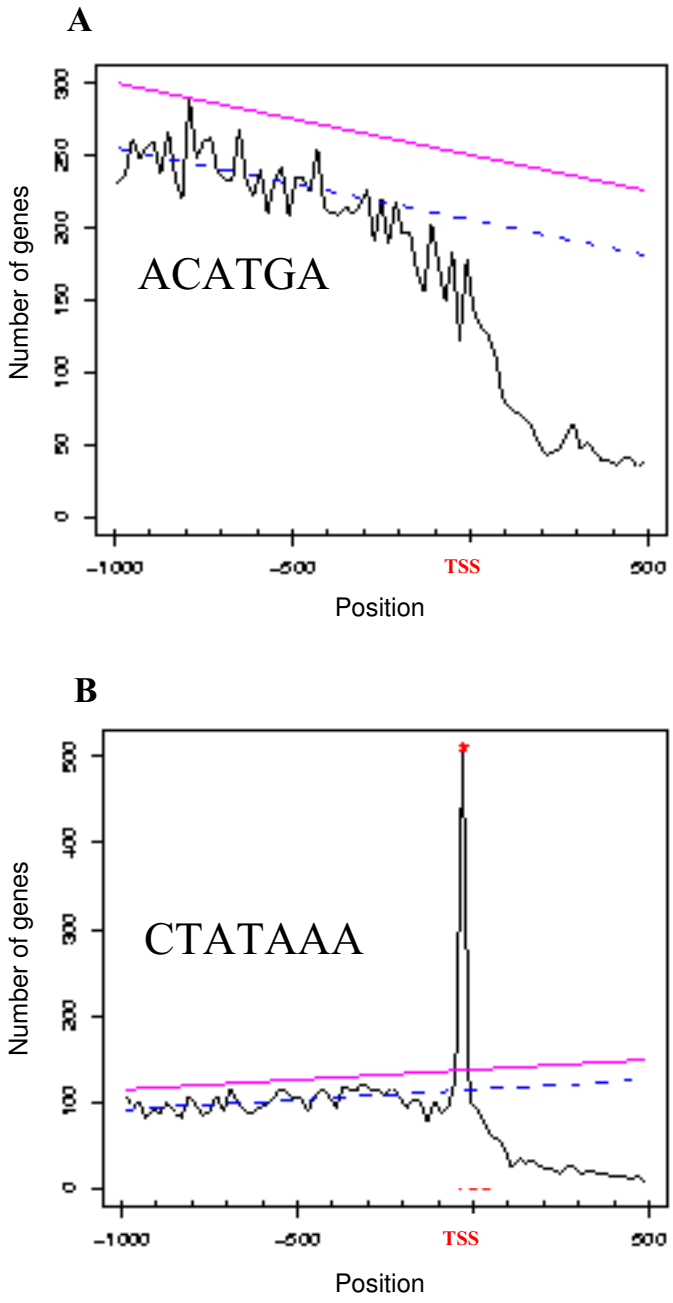

Figure 25 – *Ab initio* method for the definition of biological significant motifs and their preferential positions

Figures showing the criteria using in *ab initio* method to define biological significant motifs and their preferential positions. The motif frequencies in all the input promoters between the position -1000 and -300 from TSS are used to define an expected value (blue dotted line). A confidence interval (pink line) is then calculated and used to define the significant peaks. In the showed figures we can see how motif ACATGA (A) have no preferential position while CTATAAA motif (B) present a clear peak around -50 to -10. These figures are kindly provided by Virginie Bernard (Bernard *et al*, 2006).

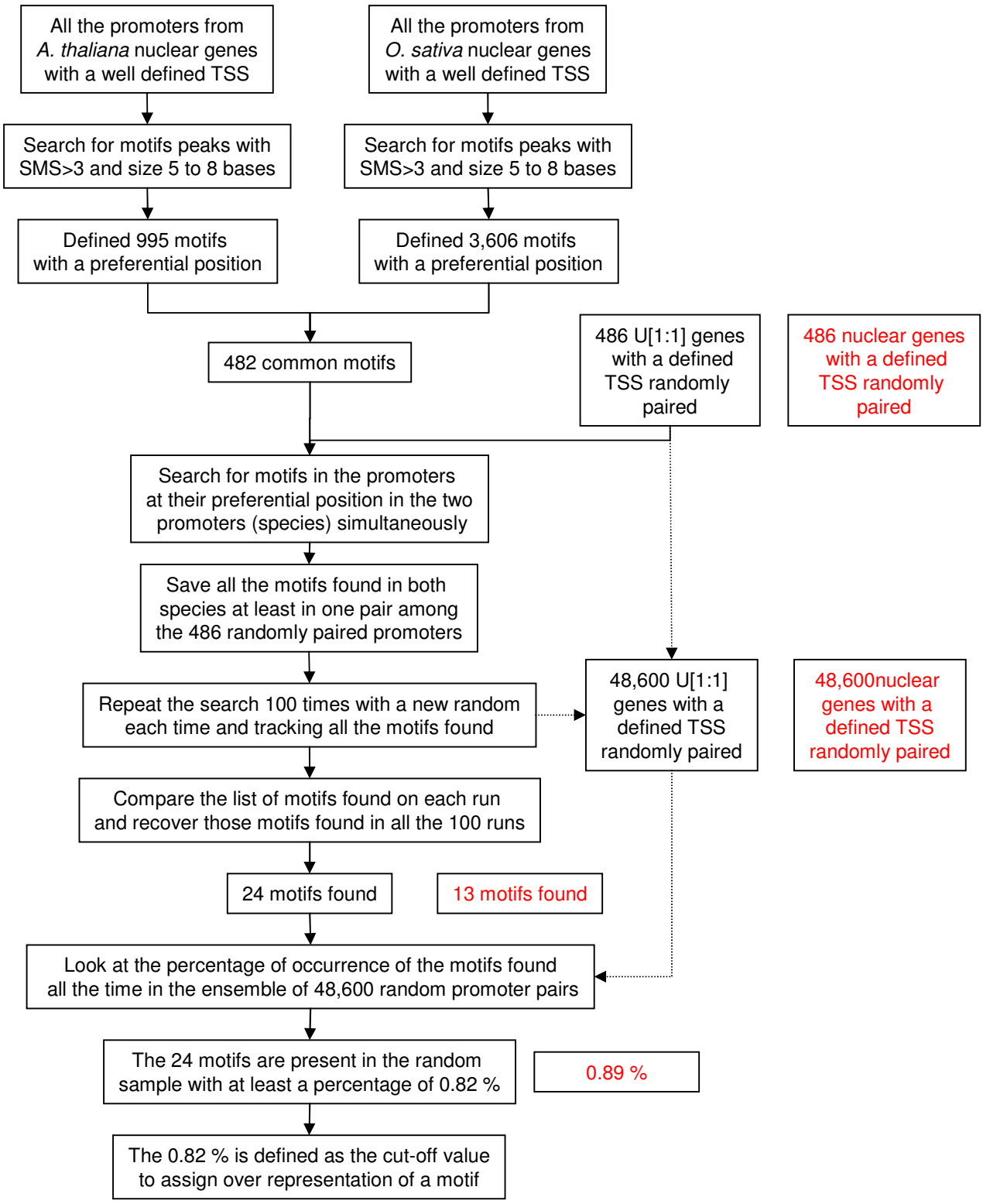

Figure 26 – Definition of criteria to detect over-represented motif
Schematic representation of the pipeline followed to define the over-representation criteria with a 'bootstrap' like analysis.

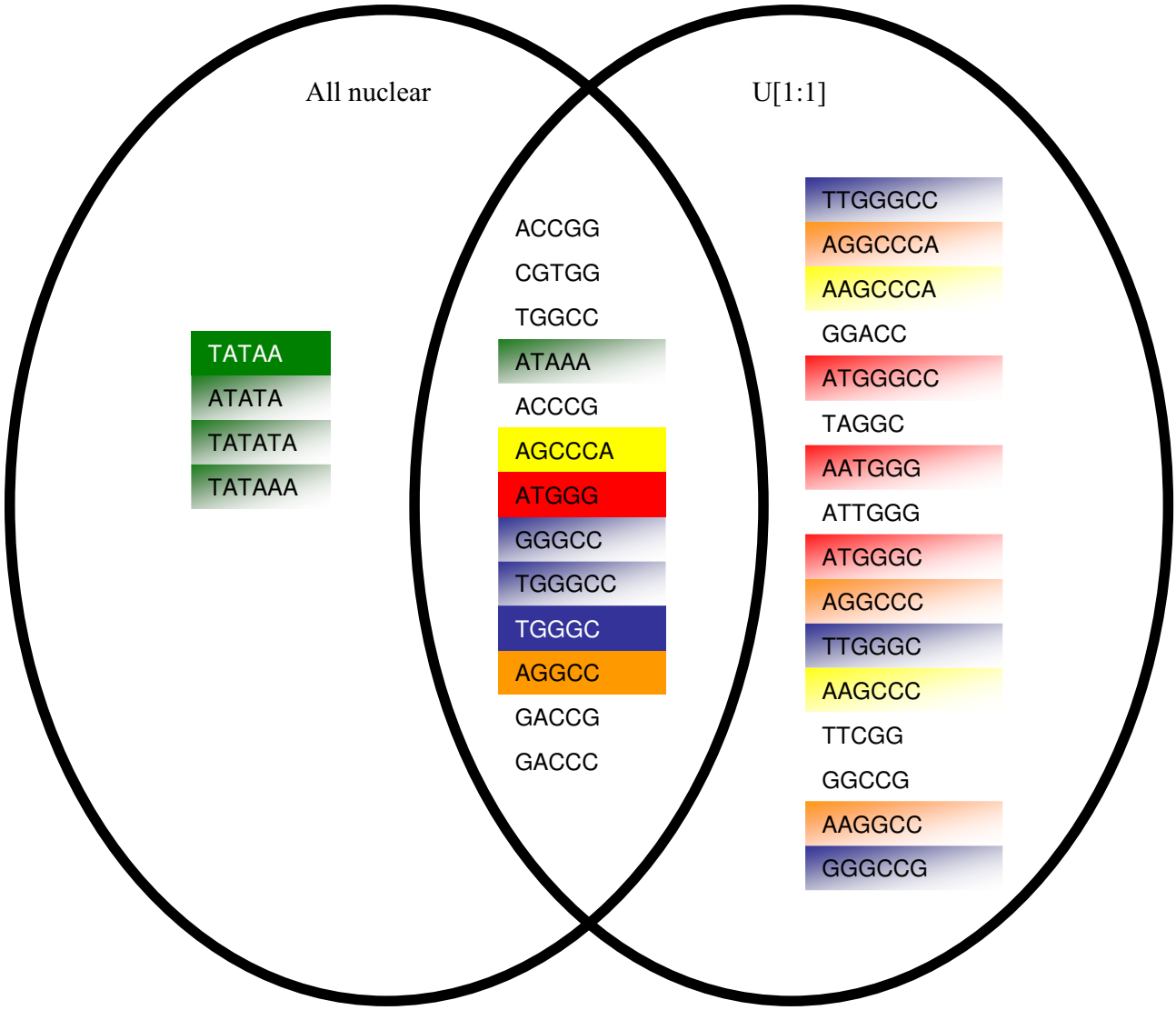

Figure 27 – Overrepresented motifs in promoters of all nuclear genes and U[1:1] genes
List of overrepresented motifs in each group of genes. Motif lists contain all the motifs of minimum size 5 nucleotides defined from analysing the coding strand of all the nuclear genes with a defined TSS. Preferential positions of significant motifs were calculated with a window of a fixed size of 20 nucleotides and used to calculate the percentage of occurrences within each group of promoters. Overrepresentation definition was based in an 'bootstrap' analysis. To highlight motif similarities, 'core' motifs are solid coloured while related 'derived' motifs are shaded. Blank motifs could not be related to any 'core' motifs.

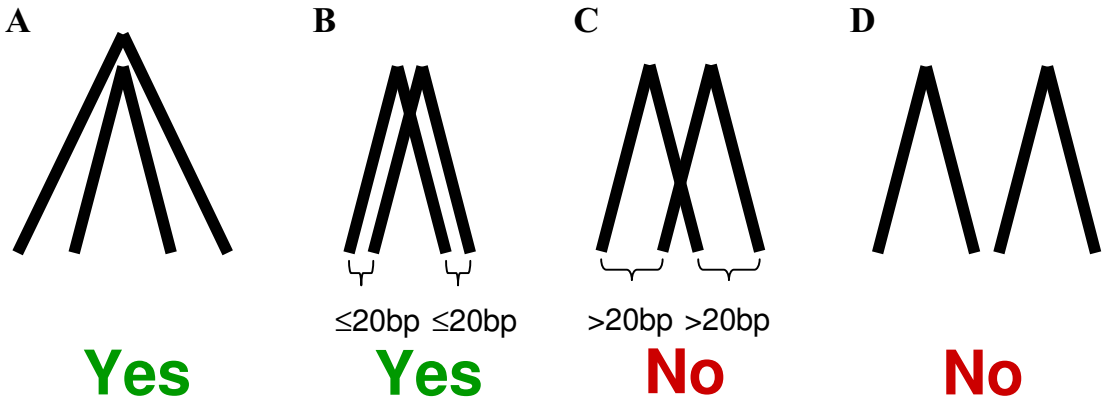

Figure 28 – Common motif peaks definition
Example of peak positions in two species defined as conserved. Peak positions in one species including peak position in second species (A) or differing up to 20 nucleotides (B) are considered as conserved, while the rest of cases are considered as not conserved (C and D).

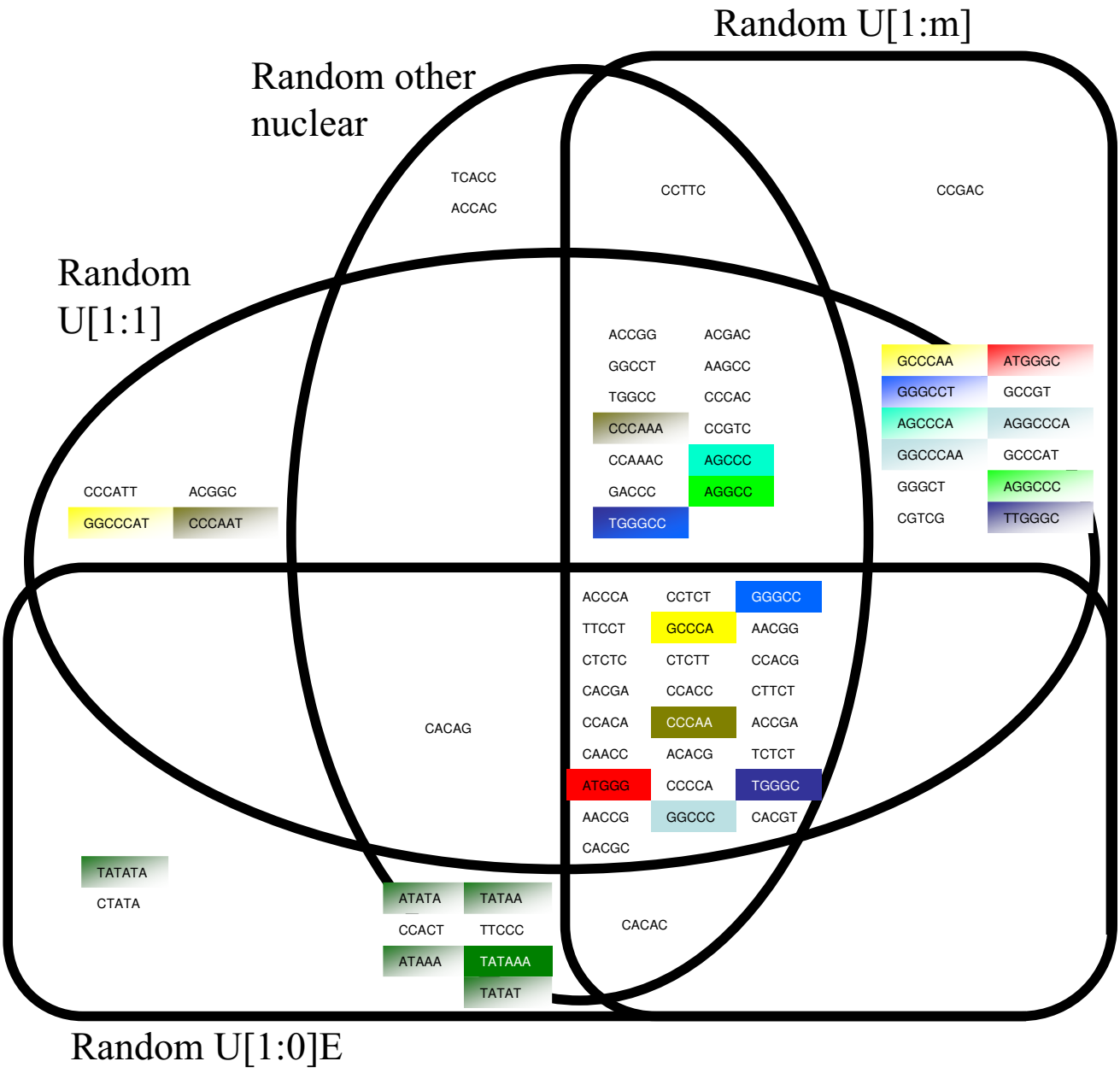

Figure 29 – Overrepresented motifs in different subgroups of promoters of unique genes
Venn diagram presenting all the 69 distinct motifs defined as overrepresented (with a conserved preferential position in *Arabidopsis thaliana* and *Oryza sativa*) in the different groups of unique genes. To highlight motif similarities, 'core' motifs are solid coloured while related 'derived' motifs are shaded. Blank motifs could not be related to any 'core' motifs.

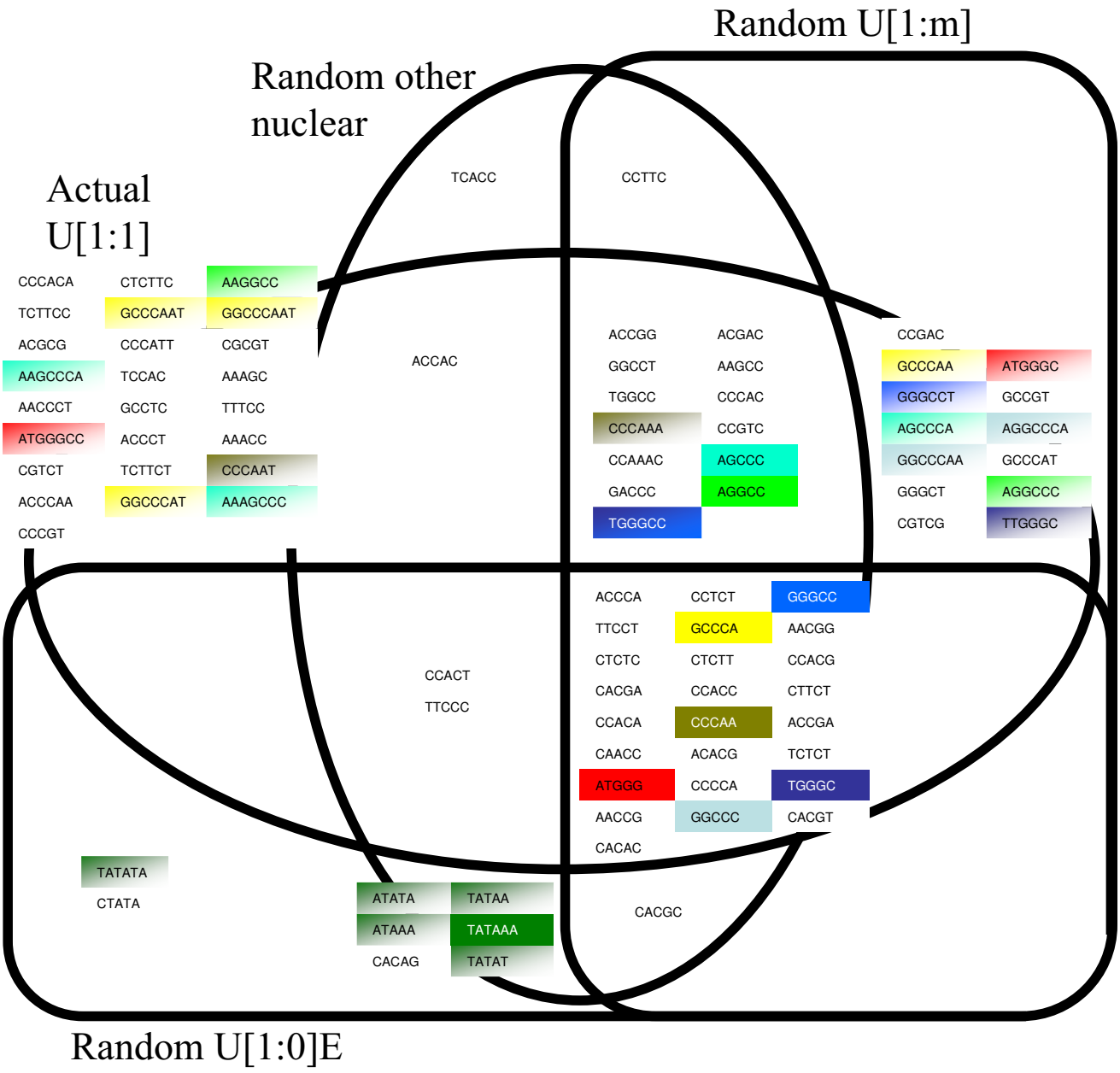

Figure 30 – Overrepresented motifs in all nuclear genes, different subgroups of unique genes (random pairs) and actual U[1:1] promoters (orthologous pairs)
To highlight motif similarities, 'core' motifs are solid coloured while related 'derived' motifs are shaded. Blank motifs could not be related to any 'core' motifs.

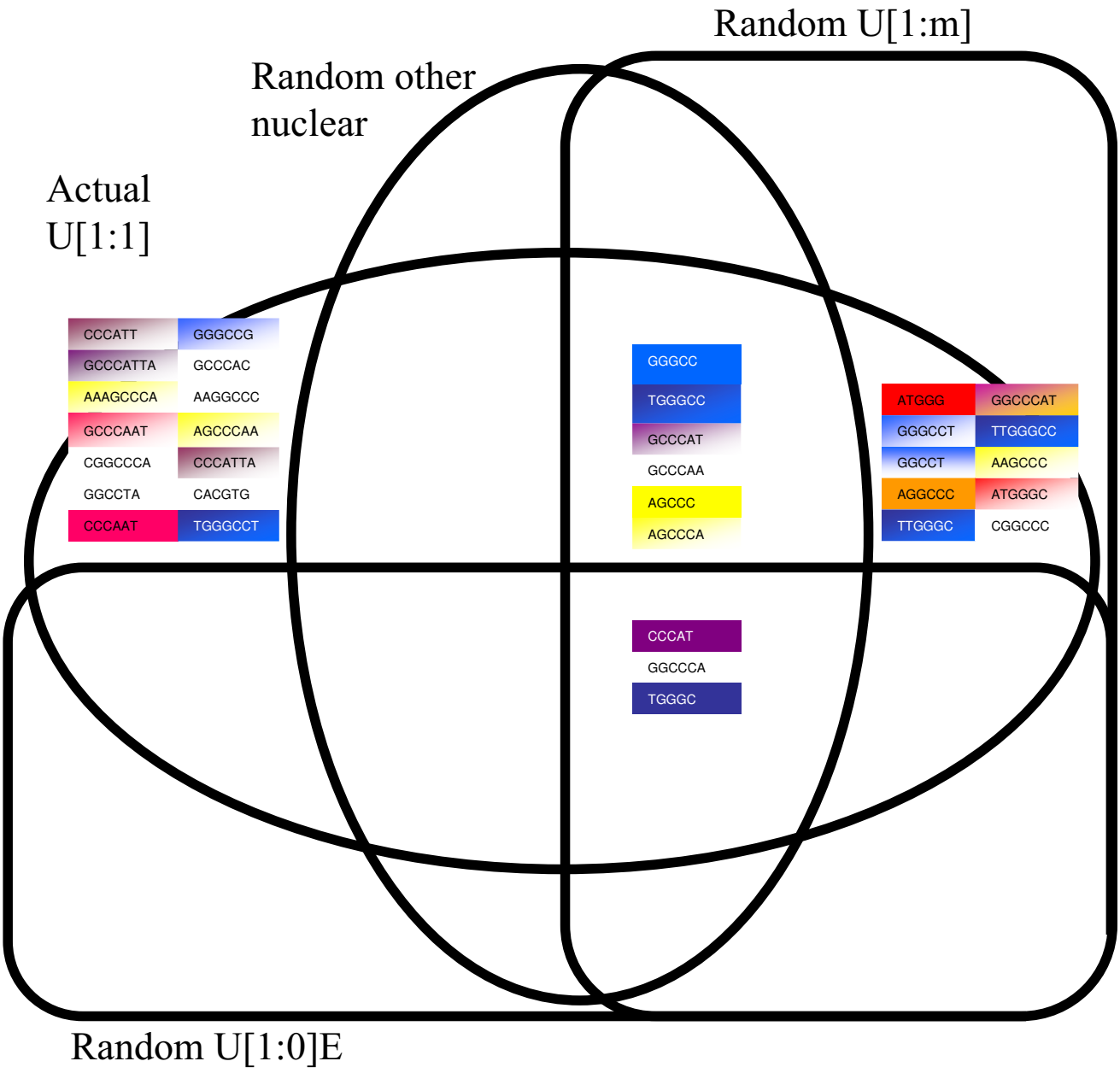

Figure 31 – Overrepresented motifs (based on U[1:1] promoters analysis) in all nuclear genes, different subgroups of unique genes (random pairs) and actual U[1:1] promoters (orthologous pairs)

To highlight motif similarities, 'core' motifs are solid coloured while related 'derived' motifs are shaded. Blank motifs could not be related to any 'core' motifs.

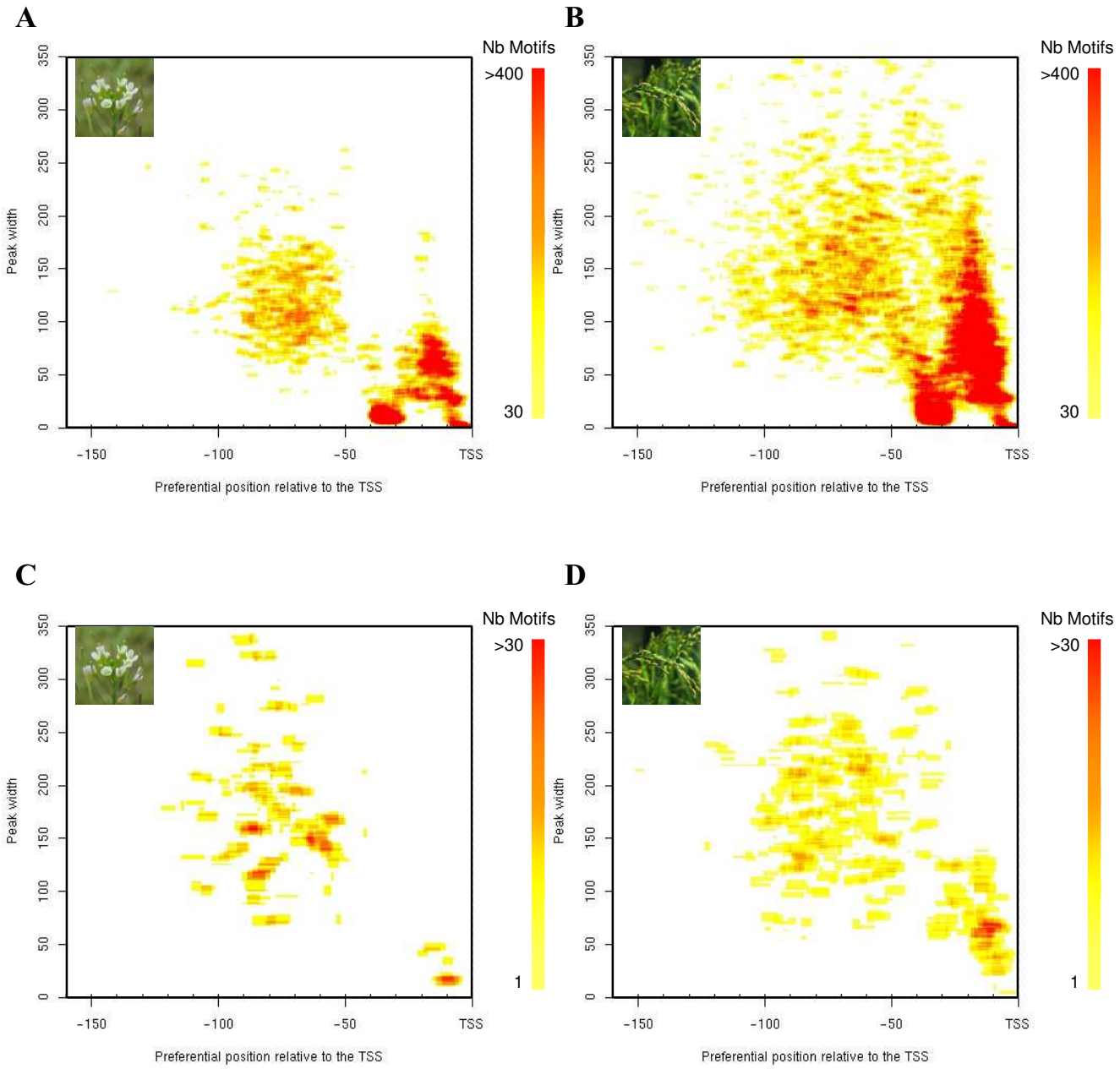

Figure 32 – Preferential positions of motifs defined from promoters of all nuclear and U[1:1] genes

Coloured representation of the preferential positions and peak width of the motifs defined from analysing the coding strand of the promoters of all the nuclear genes in *Arabidopsis thaliana* (A) and *Oryza sativa* (B) and all the promoters of AtU[1:1] (C) and OsU[1:1] (D) genes with a defined TSS.

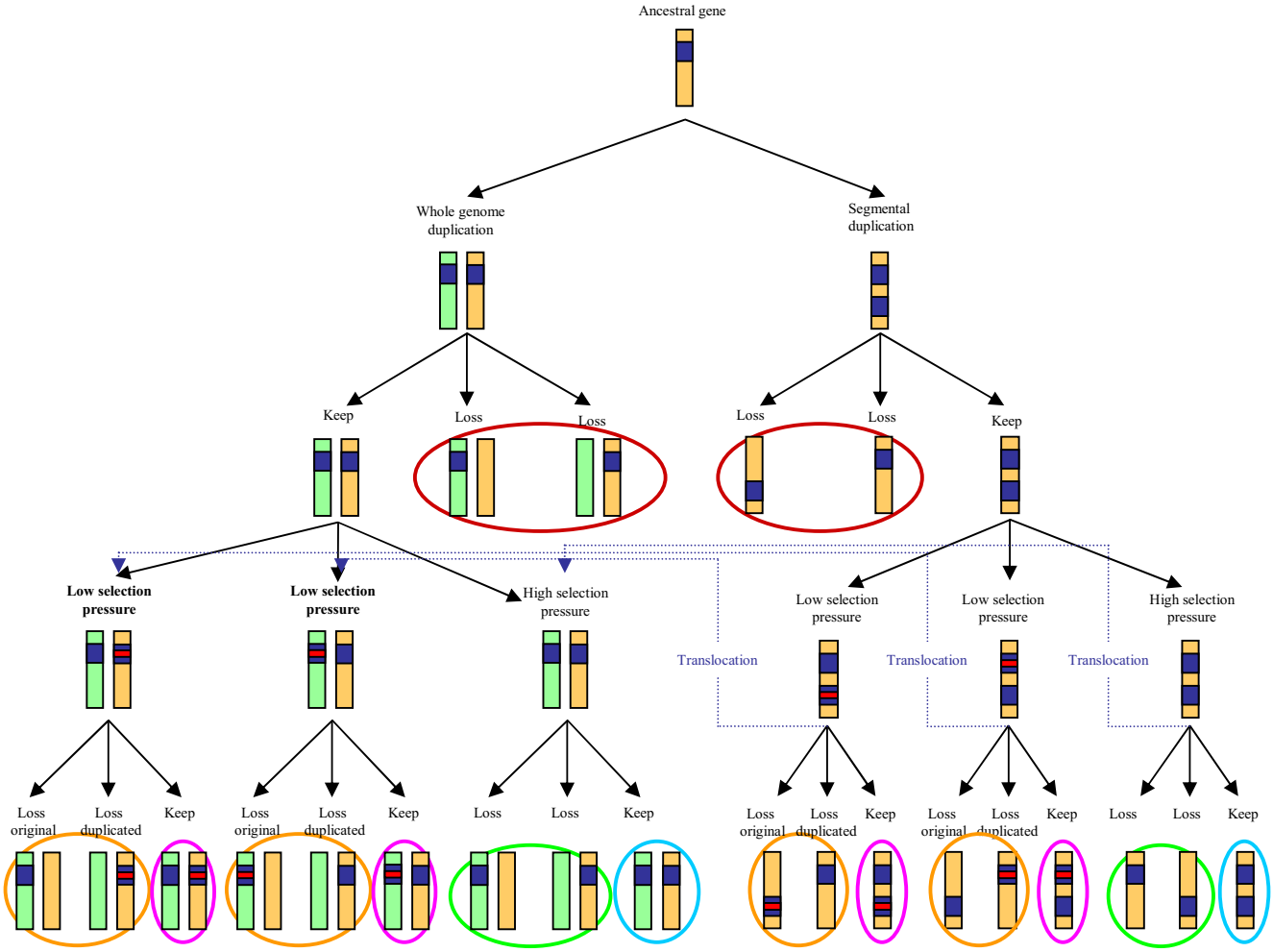

Figure 33 – Schematic possible conservation pathways followed by unique genes along evolution
Schematic illustration representing the different possible pathways followed by unique genes along evolution. Figure is centred in the possible consequences of unique genes after duplication. Fate of unique genes after duplication can be resumed in 5 different possibilities:

1. Rapid lost after duplication (Red circle): In such situation duplicated gene is exactly identical to the ancestral genes and is therefore impossible to differentiate the 'original' from the 'duplicated' after gene loss.
2. Lost after a high selection pressure period (Green circle): Like previous case except in this case the probabilities of accumulate differences is increased the longer the pressure period is. In this case, despite we would probably be not capable of differentiate 'original' from the 'duplicated' after gene loss, the remaining gene may have accumulated enough small differences to produce DMIs.
3. Lost after a low selection pressure period (Orange circle): The genes that follow this pathway accumulate mutations that may change their original functions. If the gene accumulating mutations is finally lost while the other conserve ancestral sequence we would probably qualify it as the 'original'. On the contrary, if the final unique gene is the one accumulating the mutations we would probably qualify as the 'duplicated' in a phylogeny analysis.
4. Fixation after a high selection pressure period (Blue circle): In such situation we would no longer talk about unique genes but about duplicated genes. If the selection pressure is high enough, they will accumulated small differences but not enough to distinguish the 'original' one except if we consider synteny compared with other species.
5. Fixation after a low selection pressure period (Pink circle): Finally, it is possible that unique genes are duplicated and fixed after a low selection pressure period. Like in the previous case we would not longer talk about unique genes but about duplicated genes or, even, gene family. In this case it is possible that one or both genes have accumulated mutations acquiring new functions or sub-functionalizing the ancestral ones. In this case the 'original' tag would be given to the gene accumulating less divergence when compared with ancestral one (ortholog in another species) or by synteny analyses.