# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

**4,800**
Open access books available

**122,000**
International authors and editors

**135M**
Downloads

Our authors are among the

**154**
Countries delivered to

**TOP 1%**
most cited scientists

**12.2%**
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

## Interested in publishing with us?
## Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

4

# Understanding LiP Promoters from *Phanerochaete chrysosporium*: A Bioinformatic Analysis

Sergio Lobos[1], Rubén Polanco[2], Mario Tello[3], Dan Cullen[4], Daniela Seelenfreund[1] and Rafael Vicuña[5]

[1]*Laboratorio de Bioquímica, Departamento de Bioquímica y Biología Molecular, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile*

[2]*Escuela de Bioquímica, Facultad de Ciencias Biológicas, Universidad Andrés Bello*

[3] *Centro de Biotecnología Acuícola, Universidad de Santiago de Chile.*

[4]*USDA Forest Service, Forest Products Laboratory, Madison, WI 53726,*

[5] *Departamento de Genética Molecular y Microbiología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile and Millennium Institute for Fundamental and Applied Biology*

[1,2,3,5]*Chile*

[4]*USA*

## 1. Introduction

DNA contains the coding information for the entire set of proteins produced by an organism. The specific combination of proteins synthesized varies with developmental, metabolic and environmental circumstances. This variation is generated by regulatory mechanisms that direct the production of messenger ribonucleic acid (mRNA) and subsequent translation of the nucleotide sequence into amino acid sequences, among other fundamental processes including post-translational modifications. A major step of gene expression regulation is the control of transcription initiation by RNA polymerase II. Control systems that modulate mRNA synthesis are based on the specific recognition and interaction of proteins with cognate sites on the DNA. The complex network of DNA-protein and protein-protein interactions determines the degree of transcription of a specific sequence and defines particular expression patterns. Ultimately, the outcome of this net of interactions provides the finely-tuned response to internal clues and environmental signals (Matthews, 1992).

Understanding gene expression in complex organisms such as eukaryotes is one of the most important challenges of molecular biology. One of the most fundamental and unanswered questions is whether adaptative evolution proceeds through changes in protein-coding DNA sequences or through non-coding regulatory sequences. It has been argued that morphological change occurs mainly via non-coding changes (Haygood et al., 2010). *Diptera* studies showed that *cis*-regulatory sequences that control transcription are a common source of divergent protein expression patterns and thus of phenotypic change (Wittkopp, 2006).

Also, recent analyses of the human genome suggest a distinctive role for adaptative changes both in coding and non-coding sequences. Changes in non-coding sequences appear primarily related to changes in neural development (Haygood et al., 2010).

The last decade has witnessed an explosion of studies showing that the complex regulation of gene expression is mainly modulated by the manifold interactions between transcription factors (TFs) with their corresponding transcription factors binding sites (TFBSs) on DNA (Wei & Yu, 2007). These regulatory elements are either located proximally, in sequences upstream of the transcription start site, which are generically known as promoters, or more distantly, in sequences known as enhancers or silencers. *Cis*-regulatory elements are information processing units that are embedded in genomic DNA and which regulate gene expression. Most commonly, these *cis*-regulatory elements or modules are a few dozens to several hundred base pairs long and are comprised of multiple binding sites for transcription factors. On average, a module will have binding sites for different transcription factors and for some factors, more than one site may be present (Howard & Davidson, 2004). To date, *cis*-regulatory modules of some *Drosophila* genes have been characterized at the target site level, providing an explanation of how these sequences and gene network architectures control development in early dipteran embryos (Howard & Davidson, 2004). Modules have also been denominated "motifs" by many authors, and this is the nomenclature we will use throughout this work. Knowledge of the *cis*-regulatory elements or motifs of many genes from different species may offer insight into how these sequences control the building of the diverse structures and functional adaptations found in living organisms.

As is well known, transcription involves the binding of proteins to several sites on a promoter sequence, and in eukaryotes the action of transcription factors over long distances seems to be the rule. Transcriptional outcome can be influenced by cooperative interactions of proteins between adjacent or distant sites, mainly through the formation of DNA loops, as has been described profusely in both prokaryotic and eukaryotic organisms (Han et al., 2009; Matthews, 1992; Schleif, 1992). The property of DNA to form loops enhances the regulatory properties of proteins and expands the flexibility of systems in responding to signals that evoke cellular change.

In order to understand the functional organization of a eukaryotic promoter, in this study we used the well-studied ligninolytic fungal species *Phanerochaete chrysosporium*, and examined the promoters from a selected gene family. *P. chrysosporium* has been used as a model system in numerous studies for its production of lignin-degrading enzymes (Singh & Chen, 2008). Cellulose and lignin constitute the most abundant forms of organic carbon and their degradation and mineralization is a fundamental step in the carbon cycle of the biosphere. The use of lignocellulosic biomass depends on either the removal or disruption of lignin by a process that can include the activity of lignin- and manganese-dependent peroxidases in order to expose the cellulose polymer to the attack of cellulolytic enzymes. Therefore, an understanding of the regulatory mechanisms that underlie the production of these enzymes is of pivotal importance both for a deeper comprehension of the crucial process of maintenance of the carbon cycle in nature and for the production of bioenergy. Additionally, lignocellulosic wastes are produced in large amounts and efforts have been made to convert these residues into valuable products such as biofuels, chemicals and animal feed (Dashtban et al., 2009). This bioconversion usually requires a multistep process involving a pretreatment (mechanical, chemical or biological) and hydrolysis to produce readily metabolyzable molecules such as hexoses and pentoses (Sánchez, 2009).

Pretreatment of lignocellulosic residues is necessary because hydrolysis of non-pretreated material is slow and results in low yield (Dashtban et al., 2009). It has been reported that the use of *P. chrysosporium* is advantageous for pretreatment of cotton stalks in an energy-saving, low cost and environmentally friendly approach that can reduce chemical pretreatments (Shi et al., 2009). Reported recovery depended on culture conditions, either agitated or shallow stationary submerged. Although agitated cultivation resulted in better delignification, pretreatment under submerged shallow stationary conditions provides a better balance between lignin degradation and carbohydrate availability (Shi et al., 2009). Interestingly, under solid-state cultivation, higher cellulolytic but not ligninase activity was associated with $Mn^{2+}$ addition, although the initial purpose of supplementing $Mn^{2+}$ was to improve ligninase activities and lignin degradation (Shi et al., 2008). This fungus has also shown promising results in wood biopulping (Singh et al., 2010) and soil bioremediation (Jiang et al., 2006). Hence, optimization of these biotechnological processes can also profit from a deeper understanding of the fundamental process of gene transcription.

Woodrotting fungi include white-rot basidiomycetes, brown-rot basidiomycetes, and soft-rot ascomycetes/deuteromycetes; however, only a small group of these are able to completely degrade lignin to carbon dioxide and thereby gain access to the carbohydrate polymers of plant cell walls, which they use as carbon and energy sources. Selective degradation of lignin by these fungi leaves behind crystalline cellulose with a bleached appearance that is often referred to as "white rot" (Martínez et al., 2004). Some or all of these enzymes and their isozymes of the lignin depolymerization system include multiple isozymes of lignin peroxidase (LiP) and manganese-dependent peroxidase (MnP) (Kirk & Farrell, 1987; Farrell et al., 1989; Singh & Chen, 2008). Among the ligninolytic fungi, *P. chrysosporium* is considered as a model organism for the development and understanding of the ligninolytic-enzyme-production system, as it can produce a more complete ligninolytic enzyme complex than most other species (Kirk & Farrell, 1987) and until recently, it was the only ligninolytic fungus whose genome has been sequenced (Martínez et al., 2004). In *P. chrysosporium*, LiPs together with MnPs and $H_2O_2$-producing enzymes constitute the major components of the lignin-degrading system that are secreted to the extracellular medium (Kirk & Farrell, 1987; Farrell et al., 1989; Kirk et al., 1990).

The characterization of ligninolytic enzyme systems of several basidiomycetes has revealed that in some species LiP activity is not observed. For example, in the white rot fungus *Phanerochaete sordida* only MnP activity, but no lignin peroxidase or laccase activity was detected, although several culture conditions were assayed. In this species, three highly similar MnP isoenzymes were identified (Rüttimann-Johnson et al., 1994). The white-rot basidiomycete *Ceriporiopsis subvermispora* produces two families of ligninolytic enzymes, MnPs and laccases (Lobos et al., 1994), but lignin peroxidase activity is not detected (Rajakumar et al., 1996). In *Ganoderma lucidum* low levels of MnP activity are detected in some culture media, but not in others and no LiP activity was seen in any of the media tested (D'Souza et al., 1999).

The genome of *P. chrysosporium* contains a large group of genes coding for low–redox peroxidases (LRP), including 10 *lip* genes, 5 genes coding for MnPs, 4 genes encoding multicopper copper oxidases (related to laccases) and an interesting peroxidase gene unlinked to all peroxidases, that shares residues common to both MnPs and LiPs (Martínez et al., 2004). Other white rot fungi, such as *C. subvermispora* (Rajakumar et al., 1996) and G. *lucidum* (D'Souza et al., 1999) also contain *lip*-like genes, but as described above, do not exhibit detectable LiP activity.

The recent genome sequencing of a second basidiomycete, the brown-rot fungus *Postia placenta,* yielded exciting novelties: genes encoding the class II secretory peroxidases LiP, MnP and versatile peroxidase were not detected in the *P. placenta* genome (Martínez et al., 2009). This fungus contains only one LRP gene that is not closely related to LiP and MnP, but is part of an assemblage of ''basal peroxidases'' that includes the novel peroxidase (NoP) of *P. chrysosporium* (Martínez et al., 2009). Comparison of the *P. placenta* and *P. chrysosporium* genomes indicates that the derivation of brown-rot is characterized largely by the contraction or loss of multiple gene families that are thought to be important in typical white-rot, such as cellulases, LiPs, MnPs, copper radical oxidases, among other enzymes. Phylogenetic analysis suggests that LiP and MnP gene lineages of *P. chrysosporium* were independently derived from the basal peroxidases before the divergence of *Postia* and *Phanerochaete.* If so, then the absence of LiP and MnP in *P. placenta* may reflect instances of gene loss (Martínez et al., 2009). This general pattern of simplification is consistent with the view that brown-rot fungi, having evolved novel mechanisms for initiating cellulose depolymerization, have cast off much of the energetically costly lignocellulose-degrading apparatus that is retained in white-rot fungi, such as *P. chrysosporium* (Martínez et al., 2009).
LiPs from *P. chrysosporium* are encoded by ten structurally related genes (Stewart & Cullen, 1999). The genomic organization of the *lip* genes that encode these isoenzymes is known: four genes (*lipA, lipB, lipC* and *lipE*) reside within a 35 Kb region and the remaining genes (*lipG, lipH, lipI* and *lipJ*) lie within a 15 Kb region, forming clusters where six genes occur in pairs that are transcriptionally convergent (Stewart & Cullen, 1999). The transcriptional orientation and intergenic distances indicate that regulatory promoter sequences are not shared among any of the *lip* genes. *Lip* genes have been classified by their deduced amino acid sequences and also by their intron/exon structure (Stewart & Cullen, 1999). The phylogenetic clustering defines a major subfamily I of six genes (*lipA, lipB, lipE, lipG, lipH* and *lipI*) and four minor subfamilies of only one member each (*lipC, lipD, lipF* and *lipJ*) (Stewart & Cullen, 1999).
Although the *lip* genes are structurally related and the proteins participate in a common physiological process, *lip* promoter sequences display no obvious similarities, suggesting differential gene expression of this family of isozymes. Indeed, the relative transcriptional activity of these genes has been assessed systematically, showing differential regulation in response to carbon (C)-limited or nitrogen (N)-limited culture media (Stewart & Cullen, 1999). Recently, it was shown that over a hundred proteins that are secreted by *P. chrysosporium* exhibited increased transcription in either C- or N-limited relative to nutrient replete medium, including LiP and MnP expression (Wymelenberg et al., 2009). In another study, similar expression patterns of secreted proteins between cellulose-grown and wood-grown cultures were found (Sato et al., 2007), but this study showed the complication of considering wood as a nutrient, since it is both N-limited and C-replete. In addition to enzymes which act on lignocelluloses, proteases were found, suggesting the ability to generate nitrogen (Sato et al., 2007); depletion of nitrogen triggers the onset of secondary metabolism. Metabolic switching occurs in culture after 48 hours when linear growth ceases. After 72 hours, *P. chrysosporium* has shifted to secondary metabolism, its beginning being closely related to the appearance of LiP activity (Wu & Zhang, 2010). The complex expression pattern of *lip* genes suggests that each isozyme might play a specific biological role in the process of ligninolysis, though why there is a multiplicity of lignin peroxidases remains unclear (Farrell et al., 1989; Stewart & Cullen, 1999; Sato et al., 2007). This long standing question is especially intriguing and paradoxical, since LiPs are low–redox

enzymes that catalyze a unique nonspecific enzymatic "combustion", i.e. susceptible aromatic substrate molecules are oxidized by one electron and this produces unstable cation radicals which then undergo a variety of nonenzymatic reactions (Kirk & Farrell, 1987). The answer to this fundamental issue is still a matter of debate and it is speculated that an array of different genes may provide the necessary plasticity to the fungus to attack diverse types of lignin, its recalcitrant carbon and energy source, under various biotic and abiotic conditions.

The isoenzyme family of LiP proteins from *P. chrysosporium* provides an interesting model for analyzing the evolution of promoters and their coding sequences. The identification of characteristic features regulating the main genes involved in lignin biodegradation, as well as others that are co-regulated, can both provide a more complete understanding of promoter organization and be used to identify novel genes involved in ligninolysis through bioinformatics–based searches. In this study, both bioinformatic tools and experimental data were used to explore if the structure of promoter organization is related to the phylogenetic grouping of the LiP proteins. A motif is a pattern common to a set of nucleic acid subsequences which *share some biological property of interest,* such as being a DNA binding site for a regulatory protein. It was expected that these motifs would provide information about the regulatory factors that control gene expression and identify transcription factors that bind to the motifs. The main goal was to analyze the structural organization of the promoters of the *lip* gene family and determine if there exists an organization of TFBSs and/or some kind of structured assembly of *cis*-regulatory elements or motifs within their promoter sequences. The promoter structures were compared with reported data on the differential regulation, transcription and phylogenetic analysis of the LiP proteins. To our knowledge, no reports exist where bioinformatic data has been correlated with the expression of a family of isoenzymes in filamentous fungi. The working hypothesis of this study was to establish if genes involved in the same biological process have promoters that share structural characteristics, although these common structural elements may not be evident. In this case, it should be possible detect a common architecture using appropriate bioinformatics tools in order to identify motif patterns that contain functional TFBSs.

## 2. Method and results

### 2.1 Analysis of promoters from lignin peroxidase genes

#### 2.1.1 Alignment of promoter sequences reveals a similar pattern to *lip* gene clustering

We first analyzed 1 Kb of the available promoter sequences of the ten *lip* genes using the ClustalW (Thompson et al., 1994) and the Jotun Hein algorithms (Hein, 1990) from the DNAStar software (Figure 1). Due to lack of information about the transcriptional start sites of the *lip* genes from *P. chrysosporium*, we first analyzed 1 Kb of the available region located upstream of the translational start site, since it was highly probable that promoter sequences were included. We tested two alignment algorithms: one was the Needleman-Wunsch algorithm present in ClustalW, which does not presume an evolutionary relationship between the analyzed sequences. The second was the Jotun Hein algorithm, a Markov chain algorithm that presumes an evolutionary relationship between the sequences to be analyzed. ClustalW was performed using BLOSUM62 matrix. The algorithms were used through the PC interphase provided by the DNASTART software. When using the Jotun Hein algorithm, a clustering of the promoters belonging to the subfamily I *lip* genes appeared (i.e. *lipA, lipB, lipE, lipG, lipH* and *lipI*), that is similar to the relationship between

the protein sequences (Stewart & Cullen, 1999). Cladistic analysis based on promoter sequences showed two main branches within the family. The main branch included all but one of the promoter sequences of the *lip* family I, conformed by *lipA, lipB, lipE, lipG, lipH* and *lipI*. The sole exception was *lipH*, which appeared more closely linked to the *lipF* promoter sequence. The sequences corresponding to the promoters of *lipD, lipC* and *lipJ*, which comprise the subfamilies II, III and V, respectively, were more divergent. This grouping was also apparent when the ClustalW analysis was repeated using 2 Kb of all promoter sequences. Both algorithms were able to detect an evolutionary relationship between upstream regions of the *lip* genes, but the Jotun Hein algorithm was more sensitive to detect this relation. The fact that very similar results were obtained using two different algorithms, suggests that this association is not spurious and supports the finding of a common organization of the analyzed sequences. Jotun-Hein was also used because it had been employed for analysis of the LiP proteins (Stewart & Cullen, 1999).

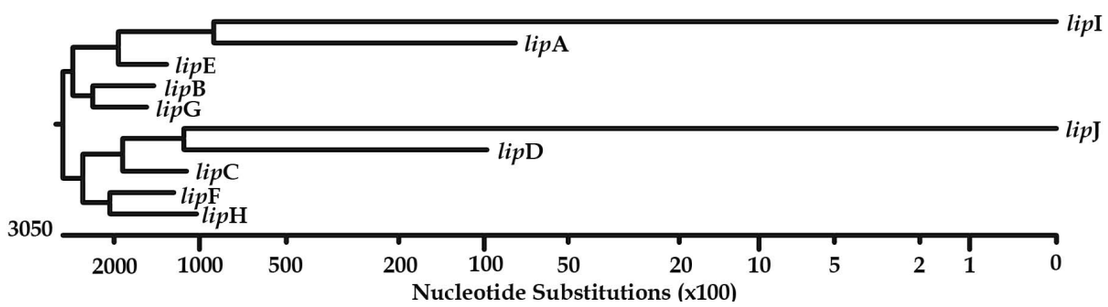



Fig. 1. Cladistic analysis of 1 Kb promoter sequences of 10 *lip* (lignin peroxidase) genes from *Phanerochaete chrysosporium.* Each sequence in the analysis corresponds to 1 Kb upstream of the ATG codon. Analysis was performed with the Jotun-Hein (Hein, 1990) algorithm on LASERGENE package software.

When the Clustal analysis included only the six promoter sequences of the genes belonging to the subfamily I of *lip* genes, a similar order appeared where *lipH* again corresponded to the most distant member of the group (Figure 2).

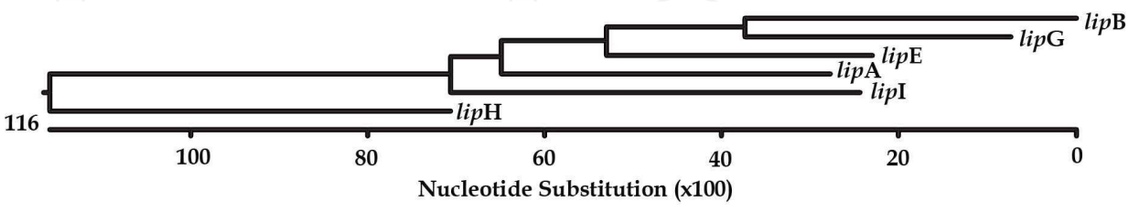



Fig. 2. Cladistic analysis of 1 Kb of six *lip* promoters corresponding to the Subfamily I classification from *Phanerochaete chrysosporium*. Each sequence in the analysis corresponds to 1 Kb upstream of the ATG codon. Analysis was done with the ClustalW (Thompson et al., 1994) algorithm on DNAStar software.

### 2.1.2 Defining an ATG upstream region for analysis

We then analyzed the available ATG upstream region of the ten *lip* genes using the Genomatix bioinformatics tool that searches conserved *cis*-regulatory elements within

TRANSFAC and JASPAR databases. It is not possible to precisely define promoter sequences, as the transcription start site is unknown in this case and it is not easy in general to define how far upstream distal sequences control gene expression. Therefore, sequences upstream of the ATG of 500, 1000 and 2000 bp were analyzed for the presence of conserved *cis*-regulatory elements or TFBSs. With this tool, a multiplicity of elements was evident; however, no clear pattern of structural organization emerged. Thus, a more sophisticated method to find sequence patterns was needed. Among programs that perform this kind of analysis, MEME (Multiple Expectation maximization for Motif Elicitation) and Gibbs are two well-documented programs for this purpose. We chose MEME because the algorithm for maximation of Multiple Expectation allows defining more clearly a motif pattern independent of its position in the sequence. On the other hand, TRANSFAC and JASPAR allow the identification of putative binding sites only for known transcription factors, but do not find new regulatory elements, especially in organisms that have not been extensively studied. When upstream sequences (500, 1000 and 2000 bp) were analyzed using MEME software, a pattern of elements emerged that split the *lip* promoters into two groups, where the genes of one group again corresponded to the members of subfamily I of *lip* genes. This separation was subtle when analyzing 500 bp or 2000 bp of the promoter sequence but was more evident when analyzing 1000 bp of the regulatory sequences. An additional reason for choosing 1 Kb ATG upstream sequences for analysis is that, as explained above, transcriptional outcome can be influenced by cooperative interactions of proteins, mainly through the formation of DNA loops. This looping depends on the probability of two sites coming together, which is optimal for cyclization at 500 bp and decreases at distances greater than 1000 bp (Matthews, 1992). For these reasons, a promoter size of 1000 bp was chosen for further studies.

### 2.2 Analytical strategy to identify regulatory elements

The next step consisted of applying a set of analytical tools to identify putative regulatory elements within the *lip* gene family. In a step-wise strategy, first putative motifs were identified with MEME; then, for each motif, integrated databases were searched for genes that contained these motifs in their promoters with the MAST software. Briefly, MAST takes any motif and transforms it into a position-dependent scoring matrix that is used to scan a curated database of promoter sequences. Finally, to identify if this sequence corresponds to a transcriptional binding site, the best match obtained in the yeast database is used by MAST to screen a transcription factors database in order to identify the TF that recognizes the yeast sequence with the motif identified by MEME. The database used for this purpose was YEASTEXTRACT. To summarize, a general streamlined approach was defined to identify a putatively functional structure in eukaryotic promoters, as outlined in Figure 3. The flowchart shows the pathway for the identification of putative motifs, TFBSs and transcription factors involved in the expression of genes containing such motifs. With this analysis, five conserved motifs were identified and characterized in the promoters of *lip* genes from *P. chrysosporium*.

#### 2.2.1 Discovery of motifs within promoters

The search for signals within the DNA sequence was carried out using MEME (Multiple EM for Motif Elicitation), a tool that was designed to discover signals (called motifs) within a set of sequences believed to share some common (but unknown) property, such as binding sites

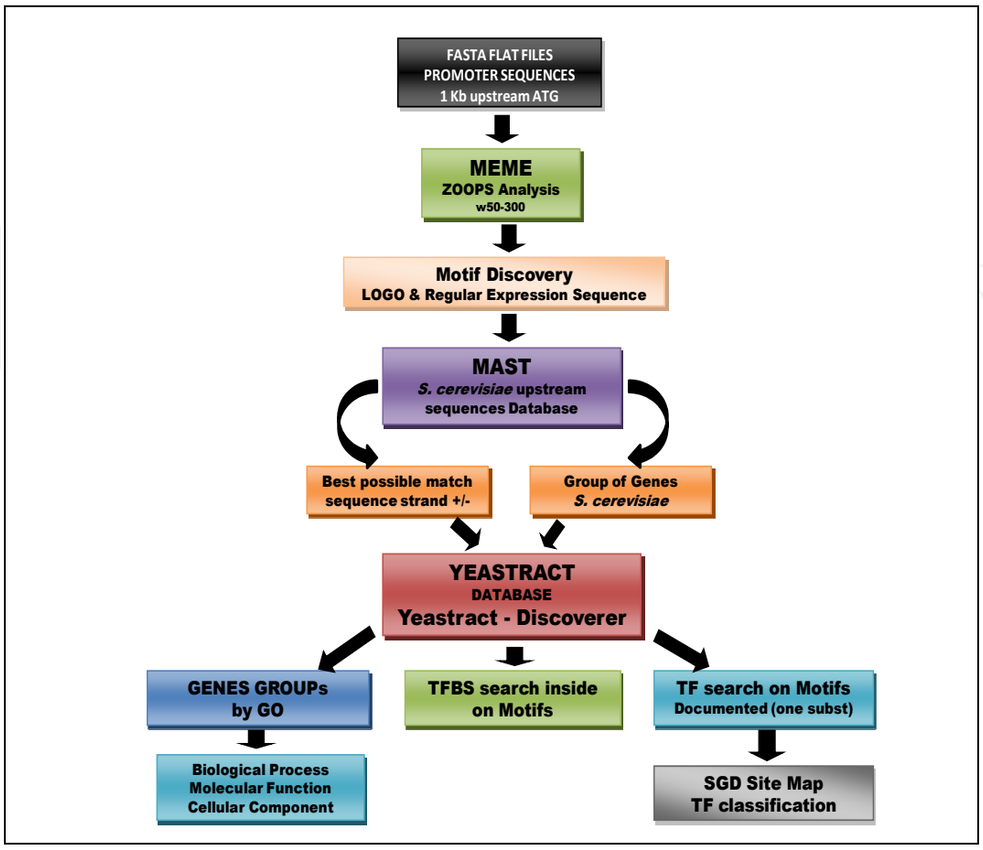



Fig. 3. Flowchart for identifying and testing putative motifs and transcription factors involved in gene expression of *lip* genes. Algorithms used at each stage are discussed in the text.

for shared transcription factors or TFBSs in a set of promoters (Bailey et al., 2006). Expectation-maximization (EM) algorithm is a method for finding maximum likelihood or maximum *a posteriori* estimates of parameters in statistical models, where the model depends on unobserved latent variables. EM is an iterative method which alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the latent variables, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step (Dempster et al., 1977). By default, MEME assumes that every position in every sequence is equally likely *a priori* to be a motif site and can search for DNA motifs on either strand (Bailey et al., 2010). MEME finds motifs by identifying highly correlated stretches of letters in the input sequences and applies statistical models to validate the most significant motifs contained in these input sequences. Finally, it reports an E-value for each motif, giving a measure of the motif's validity or likelihood of not being a random sequence artifact (Bailey & Elkan, 1994). MEME can be accessed at the web server hosted at the http://meme.ncbr.net site and is preferentially set for searching motifs within sequences of 1 Kb (Bailey et al., 2006).

A TFBS is defined as a conserved, relatively short sequence element of 10-15 bp (Stepanova et al., 2005). Since TFBSs tend to be short and degenerate, the discovery of these sequences is

a difficult task. The motif discovery algorithm searches for a minimum of two elements of similar short sequences of at least 6 bp; these motifs are searched within sliding window frames of 6 to 300 bp of width (Bailey et al., 2006). We therefore searched for motifs performing a serial analysis using 15-300, 20-300, 50-300, 100-300, 150-300, 200-300, 250-300 and 300-300 bp frames. The analysis was performed for all 10 *lip* promoters and showed a conserved pattern of motifs (Figure 4). This analysis produced a readily apparent structural organization of the *lip* genes. The motifs were most clearly noticeable with frames 15-300, 20-300 and 50-300 bp and declined with wider frames. For this reason, all further analyses were performed using the 50-300 bp window frame.

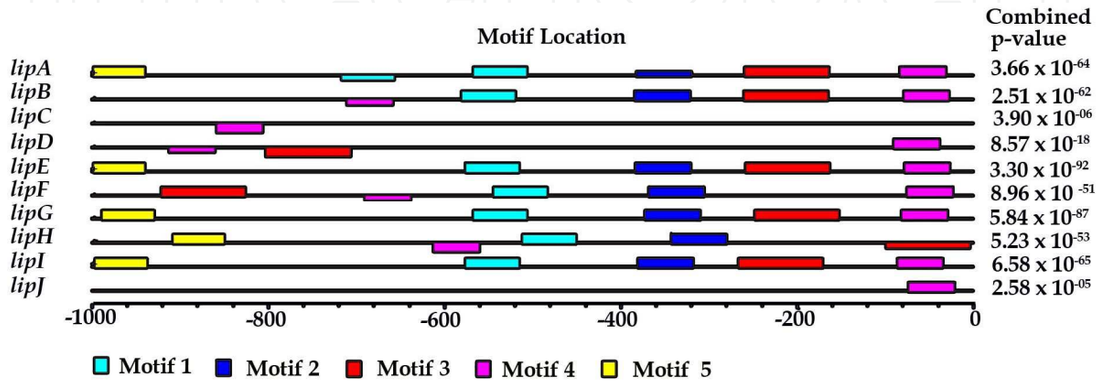



Fig. 4. The 5 most conserved motifs of the *lip* genes promoters. Maximum number of motifs: 5; windows for each motif from 50 to 300 bp. All other parameters of the MEME software corresponded to the default setting.

MEME analysis was performed using FASTA flat-files. Files containing 1 Kb of the promoter sequence from each of the 10 genes were aligned and searched for motifs with ZOOPS (Zero Or One Per Sequence) analysis. A pattern of five motifs emerged, which also corresponds to the maximum number of motifs allowed when using a 50-300 bp window frame (Figure 4). As a control, the same analysis was conducted with the promoter sequences from the subfamily I genes; when only the six promoters of the subfamily I genes were aligned, a most striking pattern of motifs emerged. Using 6-300, 20-300, 50-300 and 100-300 bp frames maintained the conspicuous pattern of five motifs that clearly indicated again a conserved organization of all six members of the subfamily I *lip* genes (Figure 5).

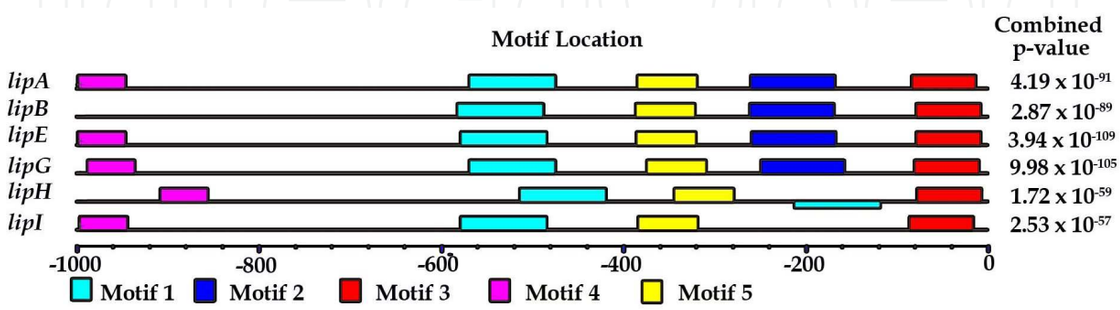



Fig. 5. Summary of the 5 most conserved motifs of the subfamily I of *lip* genes promoters. Maximum number of motifs: 5; windows per each motif from 50 to 300 bp. All other parameters of the MEME software were the default setting.

Analysis was also done using the 50-300 bp window frame or motif width window (number of characters in the sequence pattern), since the five motifs obtained presented significant E-values ranging from e-018 to e-003, which also corresponded to the best E-values of all analyzed sliding windows. Cut off E-values were set at e-003. The obtained motifs corresponded to ambiguous regular expression (LOGO) sequences of 50 to 92 bp for the five motifs identified in the subfamily I *lip* promoters and between 50 and 94 bp for the promoters of the ten *lip* genes using the MEME algorithm. To determine if the motifs found were statistically significant, the sequences were shuffled and compared to the former (training) set. Analysis of the shuffled sequences revealed that the observed motifs and the statistical significance were lost. Therefore, the structuring found in the promoter sequences was not trivial and possibly corresponds to a functional organization.

### 2.2.2 Analysis of motifs using MAST

In order to illustrate the effectiveness of the proposed strategy, as outlined in the flow sheet shown in Figure 3, analysis of the promoters of the subfamily I *lip* genes is described. The next step consisted in analyzing MEME results (LOGO) and regular expression sequences for the five motifs using MAST (Motif Alignment and Search Tool), which searches promoter motifs (best possible matches) in wide upstream sequences available in different databases. As mentioned before, MAST uses a position-dependent scoring matrix to search in a sequence for a segment with the best match. To perform this, MAST transforms any sequence pattern (motif) into a position-dependent scoring matrix. This means that a position-dependent scoring matrix is not applied to the end of a sequence or if any gap is present. The sequences are ranked according to their E-values. MAST searches databases for sequences that match the motifs and outputs detailed annotation showing genes that contain these motifs (Bailey & Gribskov, 1998) (Figure 6).

The findings of MAST in a particular upstream sequence database allowed obtaining a group of genes containing particular motifs in their promoters. The most comprehensive eukaryotic promoter databases are human, *Drosophila* and mouse; however, considering the relative phylogenetic closeness to the model fungal species *P. chrysosporium*, a yeast database was used for the analysis. The *Saccharomyces cerevisiae* genome database (SGD) is a repository of organized collection of yeast proteins and genes and their corresponding regulatory sequences and is probably the most appropriate database available today for fungal species. Using the SGD, the best possible match was found for motifs on either strand of each promoter. The obtained matches corresponded to defined and unambiguous sequences for the five motifs identified using the MEME algorithm. Sequences were subjected to MAST analysis for each separate motif and were also analyzed when combined. The best combined matches were found for these five motifs with varying E-values: motifs 1, 2 and 3 exhibited E-values of e-005, motifs 4 and 5 presented values of e-003 and e-002, respectively. The number of genes that contained the identified motifs that contained the identified motifs varied from five for motif 1, to 14 or 16 genes for motifs 2, 3 and 4 and enlarged to 23 genes for motif 5 (Table 1).

### 2.2.3 Analysis of conserved TFBSs inside each motif

Once the yeast genes that share the motif found in ATG upstream sequences of *lip* genes were obtained, transcriptional factors that bind to these sequences were analyzed with

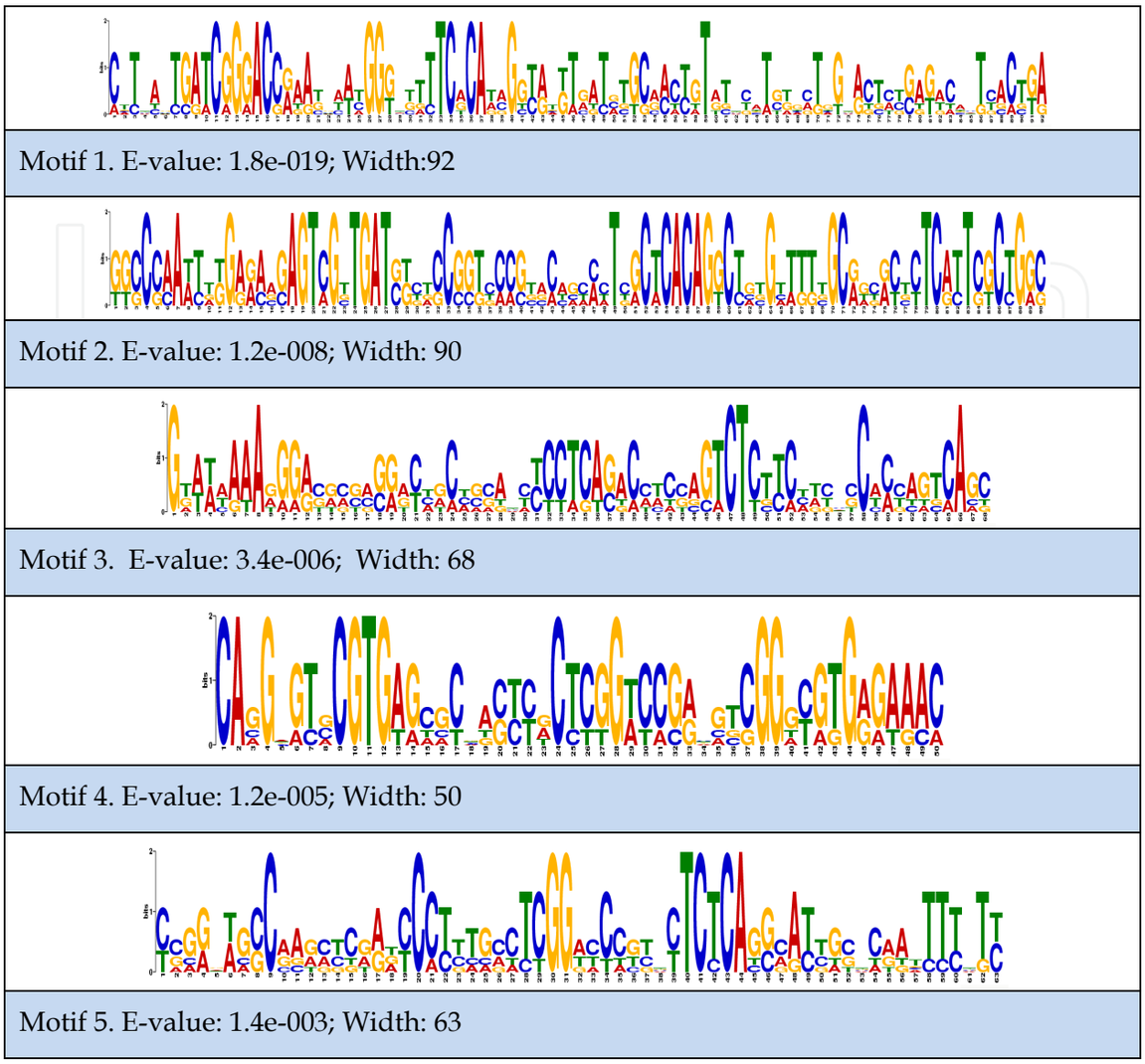



Fig. 6. LOGO representation of Motifs 1 to 5 from promoter sequences of the subfamily I *lip* genes.

YEASTRACT-DISCOVERER (YEAst Search for Transcriptional Regulators And Consensus Tracking; http://www.yeastract.com), a tool developed to support the analysis of transcription regulatory associations in yeast which can be used to identify complex motifs over-represented in promoter regions of co-regulated genes (Monteiro et al., 2008). This database contains over 48.000 documented regulatory associations between transcription factors (TFs) and target genes (Abdulrehman et al., 2011), and includes 284 specific DNA binding sites for 108 characterized TFs (Monteiro et al., 2008). To identify TFBS inside of the motifs, Yeastract uses the Smith–Waterman algorithm that allows local alignments between sequences (Smith & Waterman, 1981). The pattern matching method of YEASTRACT in search of TFBSs leads to the identification of putative target genes for specific TFs (Monteiro et al., 2008). The SGD database was therefore used to find yeast genes containing the motifs identified within the promoter sequences of the six genes of the subfamily I *lip* genes from *P. chrysosporium*.

For each motif (1 to 5), a list of yeast genes was identified. The genes that contained one of these motifs in their promoter sequence and that are included in the YEASTRACT database

were further analyzed. Each gene was queried using the SGD and finally searched with GO (Gene Onthology) and its nature determined according to three defined categories: Biological process, molecular function and cellular component (Table 2).

| Motifs 1-5 Subfamily I Group of Genes | | | | |
|---|---|---|---|---|
| Motif 1 (+) \| (-)<br>Evalue: 1e-05 | Motif 2 (+) \| (-)<br>Evalue: 1e-05 | Motif 3 (+) \| (-)<br>Evalue: 1e-05 | Motif 4 (+) \| (-)<br>Evalue: 1e-03 | Motif 5 (+) \| (-)<br>Evalue: 1e-02 |
| **YGR209C \| TRX2**<br>**YLR173W \| YLR173W**<br>**YPR094W \| RDS3**<br>**YLR246W \| ERF2**<br>**YIR042C \| YIR042C** | **YHR135C \| YCK1**<br>**YPL069C \| BTS1**<br>**YDR101C \| ARX1**<br>**YKL085W \| MDH1**<br>**YGL047W \| ALG13**<br>**YER039C-A \| YER039C-A**<br>**YER040W \| GLN3**<br>**YGL186C \| TPN1**<br>**YIL071C \| PCI8**<br>**YER038C \| KRE29**<br>**YJR148W \| BAT2**<br>**YBL066C \| SEF1**<br>**YKL065C \| YET1**<br>**YKL064W \| MNR2** | **YOL132W \| GAS4**<br>**YOR034C-A \| YOR034C-A**<br>**YDR477W \| SNF1**<br>**YPR181C \| SEC23**<br>**YPR182W \| SMX3**<br>**YDR059C \| UBC5**<br>**YDR060W \| MAK21**<br>**YHR112C \| YHR112C**<br>**YHR113W \| YHR113W**<br>**YOL162W \| YOL162W**<br>**YDR316W \| OMS1**<br>**YGR094W \| VAS1**<br>**YNL268W \| LYP1**<br>**YOR042W \| CUE5**<br>**YMR303C \| ADH2**<br>**YMR304W \| UBP15** | **YLR194C \| YLR194C**<br>**YJL006C \| CTK2**<br>**YDL174C \| DLD1**<br>**YDL173W \| YDL173W**<br>**YKL020C \| SPT23**<br>**YKL019W \| RAM2**<br>**TS(CGA)C \| SUP61**<br>**YML123C \| PHO84**<br>**YML121W \| GTR1**<br>**YML028W \| TSA1**<br>**YBR170C \| NPL4**<br>**YNL033W \| YNL033W**<br>**YBR268W \| MRPL37**<br>**YLR297W \| YLR297W** | **YFR021W \| ATG18**<br>**YGR098C \| ESP1**<br>**YCR084C \| TUP1**<br>**YCR086W \| CSM1**<br>**YKR024C \| DBP7**<br>**YKR025W \| RPC37**<br>**YLR327C \| TMA10**<br>**TS(GCU)L \| TS(GCU)L**<br>**TG(GCC)P2 \| TG(GCC)P2**<br>**TY(GUA)F2 \| SUP6**<br>**YFR028C \| CDC14**<br>**YPL103C \| FMP30**<br>**YPL101W \| ELP4**<br>**YOR180C \| DCI1**<br>**YOR181W \| LAS17**<br>**YDR178W \| SDH4**<br>**YOR140W \| SFL1**<br>**YDL079C \| MRK1**<br>**YCR076C \| YCR076C**<br>**YKL041W \| VPS24**<br>**TY(GUA)M1 \| SUP5**<br>**YBR068C \| BAP2**<br>**YGR274C \| TAF1** |
| 5 | 14 | 16 | 14 | 23 |

Table 1. Group of genes found in *S. cerevisiae* that share TFBSs found in Motifs 1 to 5.

The information of "Cellular component" for each gene was retrieved directly from the SGD database for every individual gene identified in the previous step. YEASTRACT simultaneously searches for TFBSs contained in each motif found and also searches for documented TFs that bind to these motifs (See Figure 3). This approach reduces output to a tractable size, amenable to different kinds of analysis (Table 2).

Putative functions of the identified genes suggest an interesting grouping: Motif 1 includes a single gene (Trx2) involved in cellular response to oxidative stress that presents electron carrier activity. It is noteworthy that the gene Trx2 corresponds to a cytoplasmic thioredoxin isoenzyme that is present in fungal cell walls. Motif 2 is found mainly in genes related to nitrogen metabolism and protein biosynthesis and appears to participate in biological processes of cell aging. Several of these genes are involved in biosynthetic processes of amino acids, amines and isoprenoids and also in the catabolism of amino acids. Motif 3 seems to be related to biological processes of cellular response to nitrogen and carbon metabolism and possibly, growth and differentiation. Genes containing this motif are involved in catabolic processes and cell aging, including cellular response to nitrogen starvation and eventually fungal cell wall assembly. Motif 3 is the most proximal motif identified and includes the TATA-box. This *cis*-element is conserved in all members of the subfamily I lip genes and also in all members of the *lip* gene family (in Figure 4 it

corresponds to motif 4, the most proximal regulatory element for all genes, with the exception of *lip*C). Indeed, the TATA–box is conserved in approximately 30% of all eukaryotic genes (Mariño-Ramírez et al., 2004) and therefore might correspond to an ancestral regulatory feature. TATA element recognition has remained constant over the course of evolution. Genes encoding TATA-binding proteins (TBPs) have been cloned from organisms ranging from archaea to human and all share a phylogenetically conserved 180-residue carboxyterminal or core segment, which supports all of the protein's biochemically important functions in RNA Polymerase II transcription (Patikoglou et al., 1999). Motif 4 is present in several genes that do not seem to relate to a common biological process. However, one of these is an ion transporter. The finding of this motif in a gene coding for a manganese/phosphate transporter is specially striking, since MnPs also participate in lignin catabolism. This motif corresponds to the most distal element in the studied promoter (see Figure 5). Motif 5 is related to mitosis, cell cycle, chromosome segregation and stress response. The relevance of genes associated with each motif will be discussed below.

Since all genes analyzed were identified in the yeast database (SGD), an important consideration was to determine if orthologous genes exist in the genome of *P. chrysosporium*. A preliminary search in the genome of this basidiomicete (Martínez et al., 2004) indicated that all genes shown in Table 2, with the exceptions of ARX1, MDH1, SPT23, GTR1 and DLD1, are present in the *P. chrysosporium* genome.

| MEME | MAST AND YEASTRACT ANALYSIS | *Saccharomyces cerevisiae* Genome Database | | |
|---|---|---|---|---|
| | | GO ANNOTATIONS | | |
| | | BIOLOGICAL PROCESS | MOLECULAR FUNCTION | CELLULAR COMPONENT |
| MOTIF 1 | YGR209C \| TRX2 | Cell redox homeostasis - cellular response to oxidative stress<br>ER to Golgi vesicle-mediated transport<br>Retrograde vesicle-mediated transport, Golgi to ER<br>Vacuole fusion (non-autophagic) - vacuole inheritance<br>Protein deglutathionylation<br>Regulation of DNA replication<br>Sulfate assimilation | Disulfide oxidoreductase | Cytosol<br>Fungal-type vacuole |
| MOTIF 2 | YJR148W \| BAT2 | Branched chain family amino acid biosynthetic process<br>Branched chain family amino acid catabolic process | Branched-chain-amino-acid transaminase | No manually curated |
| | YPL069C \| BTS1 | Terpenoid biosynthetic process | Farnesyltranstra-nsferase | No manually curated |
| | YKL085W \| MDH1 | Aerobic respiration - Tricarboxylic acid cycle<br>Chronological cell aging - Replicative cell aging | L-malate dehydrogenase | Mitochondrion - mitochondrial matrix |
| | YDR101C \| ARX1 | Ribosomal large subunit biogénesis | Unknown | Cytoplasm - Colocalizes with cytosolic large ribosomal subunit – Nucleoplasm |
| | YBL066C \| SEF1 | Unknown | Unknown | Unknown |
| | YER040W \| GLN3 | Nitrogen catabolite activation of transcription | Sequence-specific DNA binding transcription factor | Cytosol – Nucleus |

| | | | | |
|---|---|---|---|---|
| MOTIF 3 | YMR303C \| ADH2 | Amino acid catabolic process to alcohol via Ehrlich pathway<br>Ethanol metabolic process – Fermentation<br>NADH oxidation | Alcohol dehydrogenase (NAD) activity | No manually curated |
| | YDR477W \| SNF1 | Cellular response to nitrogen starvation<br>Invasive growth in response to glucose limitation<br>Pseudohyphal growth<br>Regulation of carbohydrate metabolic process<br>Positive regulation of gluconeogenesis<br>Biofilm formation - Cell adhesion<br>Protein phosphorylation - Signal transduction<br>Negative regulation of translation<br>Replicative cell aging | AMP-activated protein kinase activity | AMP-activated protein kinase complex<br>Cytoplasm<br>Fungal-type vacuole<br>Nuclear envelope lumen<br>Nucleus |
| | YPR182W \| SMX3 | Nuclear mRNA splicing, via spliceosome | Unknown | U1 snRNP<br>U4/U6 x U5 tri-snRNP complex<br>U5 snRNP |
| | YGR094W \| VAS1 | Valyl-tRNA aminoacylation | Valine-tRNA ligase activity | Cytoplasm<br>Mitochondrion |
| | YOL132W \| GAS4 | Ascospore wall assembly | 1,3-Beta-glucanosyltransferase | Fungal-type cell wall |
| MOTIF 4 | YKL020C \| SPT23 | Fatty acid metabolic process<br>Positive regulation of transcription from RNA polymerase II promoter<br>Response to cold | Transcription activator activity | Integral to endoplasmic reticulum membrane<br>Nucleus |
| | YML121W \| GTR1 | Chromatin silencing at telomere<br>Phosphate transport<br>Transcription from RNA polymerase I / III promoters | GDP / GTP binding | Cytoplasm -GSE complex<br>Late endosome membrane<br>Nucleus - Vacuolar membrane |
| | YBR268W \| MRPL37 | Mitochondrial translation | Structural constituent of ribosome | Mitochondrial large ribosomal subunit |
| | YDL174C \| DLD1 | Aerobic respiration<br>Cellular carbohydrate metabolic process | D-lactate dehydrogenase (cytochrome) activity | Mitochondrial inner membrane<br>Mitochondrion |
| | YML123C \| PHO84 | Manganese ion transport<br>Phosphate transport - Polyphosphate metabolic process | Inorganic phosphate transmembrane transporter activity<br>Manganese ion transmembrane transporter activity | Integral to plasma membrane |
| MOTIF 5 | YFR028C \| CDC14 | Mitotic cell cycle<br>Nucleolus organization<br>Protein dephosphorylation<br>Regulation of exit from mitosis | Phosphoprotein phosphatase | Nucleolus<br>RENT complex<br>Spindle pole body |
| | YGR098C \| ESP1 | Apoptosis<br>Mitotic sister chromatid segregation<br>Negative regulation of protein phosphatase type 2A activity<br>Regulation of exit from mitosis<br>Regulation of mitotic spindle elongation | Cysteine-type endopeptidase | Cytoplasm<br>Nucleus<br>Spindle |
| | YDR178W \| SDH4 | Cellular respiration<br>Mitochondrial electron transport, | Contributes to succinate | Mitochondrial respiratory chain |

| | | | | |
|---|---|---|---|---|
| MOTIF 5 | | succinate to ubiquinone<br>Tricarboxylic acid cycle | dehydrogenase (ubiquinone) activity | complex II |
| | YOR180C \| DCI1 | Not clearly defined | Not clearly defined | Peroxisomal matrix |
| | YOR181W \| LAS17 | Actin cortical patch localization<br>Actin filament organization / actin polymerization or depolymerization<br>Positive regulation of actin filament bundle assembly<br>Bipolar cellular bud site selection<br>Cytokinesis – Endocytosis<br>Response to osmotic stress | Cytoskeletal protein binding | Actin cortical patch |
| | YOR140W \| SFL1 | Negative regulation of transcription from RNA polymerase II promoter | Specific transcriptional repressor activity<br>Specific RNA polymerase II transcription factor activity<br>Transcription activator activity | Nuclear chromosome |
| | YKR024C \| DBP7 | Ribosomal large subunit assembly<br>rRNA processing | ATP-dependent RNA helicase activity | Nucleolus |
| | YDL079C \| MRK1 | Protein phosphorylation<br>Regulation of protein catabolic process<br>Response to stress | Protein serine/threonine kinase activity | Unknown |
| | YGR274C \| TAF1 | Gene-specific transcription from RNA polymerase II promoter<br>General transcription from RNA polymerase II promoter<br>RNA polymerase II transcriptional preinitiation complex assembly | Chromatin binding<br>RNA polymerase II transcription factor activity<br>Histone acetyltransferase activity<br>Protein complex scaffold<br>TATA-binding protein binding | Transcription factor TFIID complex |
| | YKL041W \| VPS24 | Intralumenal vesicle formation<br>Late endosome to vacuole transport<br>Ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway | Protein binding | Cytoplasm<br>ESCRT III complex |

Table 2. List of relevant genes obtained by YEASTRACT and grouped by motif. GO classification is described for each gene.

### 2.2.4 Search of transcription factors that recognize TFBSs inside motifs

The YEASTRACT database also makes publicly available up-to-date information on documented regulatory associations between TFs and DNA-binding sites in *S. cerevisiae*. Information in this database has been curated on precise tests of the associations between TFs and DNA-binding sites provided by experiments such as Chromatin ImmunoPrecipitation (ChIP), ChIP-on-chip and Electrophoretic Mobility Shift Assay (EMSA), that prove the direct binding of the TF to the target gene promoter region. Alternatively, the effect on target-gene expression of the site-directed mutation of the TF binding site in its promoter region was also considered by direct experimental evidence, which strongly suggests that the TF interacts with that specific target (Abdulrehman et al.,

2011). Analysis of TFs that bind to TFBSs from genes listed in Table 2 was performed for motifs 1 to 5. The TFs found to bind to motif 1 are shown in Table 3. The identified TFs are mainly involved in the control of the cell cycle and unfolded protein response, and to a lesser extent, in inter-organelle communication and energy metabolism. TFs that recognize motifs 2-5 also include Ash1p, Hac1p and Mot3p. Strikingly, the transcription factor Stb5p, an activator of multidrug resistance genes, binds to motifs 2, 4 and 5. Other TFs identified are also involved in the regulation of energy metabolism and cell cycle. It is important to point out that single base changes in the tested TFBSs dramatically increase the number of putative TFs that bind to them, suggesting that the identified TFs are not likely to be chosen randomly.

| Transcription Factor | Consensus | Position | Strand | Protein Info | |
|---|---|---|---|---|---|
| **Target Sequence: Motif1_Subfamily I (size 92)** | | | | | |
| Ace2p, Swi5p | ACCAGC | -19 | R | Transcription factor that activates expression of early G1-specific genes, localizes to daughter cell nuclei after cytokinesis and delays G1 progression in daughters, localization is regulated by phosphorylation; potential Cdc28p substrate | Transcription factor that activates transcription of genes expressed at the M/G1 phase boundary and in G1 phase; localization to the nucleus occurs during G1 and appears to be regulated by phosphorylation by Cdc28p kinase |
| Ash1p | YTGAT | -87, -48 | F | Zinc-finger inhibitor of HO transcription; mRNA is localized and translated in the distal tip of anaphase cells, resulting in accumulation of Ash1p in daughter cell nuclei and inhibition of HO expression; potential Cdc28p substrate | |
| | | -68 | R | | |
| Hac1p | CCAGC | -20 | R | bZIP transcription factor (ATF/CREB1 homolog) that regulates the unfolded protein response, via UPRE binding, and membrane biogenesis; ER stress induced splicing pathway utilizing Ire1p, Trl1p and Ada5p facilitates efficient Hac1p synthesis | |
| Mot3p | WAGGTA<br>TAGGTA | -55 | F | Nuclear transcription factor with two Cys2-His2 zinc fingers; involved in repression of a subset of hypoxic genes by Rox1p, repression of several DAN/TIR genes during aerobic growth, and repression of ergosterol biosynthetic genes | |

| | | | | | |
|---|---|---|---|---|---|
| Rtg1p, Rtg3p | GGTAC | -22 | F | Transcription factor (bHLH) involved in interorganelle communication between mitochondria, peroxisomes, and nucleus | Basic helix-loop-helix-leucine zipper (bHLH/Zip) transcription factor that forms a complex with another bHLH/Zip protein, Rtg1p, to activate the retrograde (RTG) and TOR pathways |
| **Target Sequence: Motif2_Subfamily I (size 90)** | | | | | |
| Ash1p | YTGAT | -68 | F | Zinc-finger inhibitor of HO transcription; mRNA is localized and translated in the distal tip of anaphase cells, resulting in accumulation of Ash1p in daughter cell nuclei and inhibition of HO expression; potential Cdc28p substrate | |
| Hac1p | CCAGC | -1 | R | bZIP transcription factor (ATF/CREB1 homolog) that regulates the unfolded protein response, via UPRE binding, and membrane biogenesis; ER stressinduced splicing pathway utilizing Ire1p, Trl1p and Ada5p facilitates efficient Hac1p synthesis | |
| Stb5p | CGGNS | -58 | F | Activator of multidrug resistance genes, forms a heterodimer with Pdr1p; contains a Zn(II)2Cys6 zinc finger domain that interacts with a PDRE (pleotropic drug resistance element) in vitro; binds Sin3p in a two-hybrid assay | |
| | | -56 | R | | |
| **Target Sequence: Motif3_Subfamily I (size 68)** | | | | | |
| Mot3p | AAGGKA | -62 | F | Described before | |
| **Target Sequence: Motif4_Subfamily I (size 50)** | | | | | |
| Stb5p | CGGNS | -25, -14 | F | Described before | |
| | | -18, -40 | R | | |
| Gsm1p | CGGNNNNNNNNCGG | -25 | F | Putative zinc cluster protein of unknown function; proposed to be involved in the regulation of energy metabolism, based on patterns of expression and sequence analysis | |
| Target Sequence: Motif5_SubfamilyI (size 63) | | | | | |
| Gcr1p | CTTCC | -56 | R | Transcriptional activator of genes involved in glycolysis; DNA-binding protein that interacts and functions with the transcriptional activator Gcr2p | |

| | | | | |
|---|---|---|---|---|
| Mot3p | CAGGYA | -21 | F | Described before |
| Stb5p | CGGNS | -35 | F | Described before |
| Xbp1p | CTCGA | -51 | F | Transcriptional repressor that binds to promoter sequences of the cyclin genes, CYS3, and SMF2; expression is induced by stress or starvation during mitosis, and late in meiosis; member of the Swi4p/Mbp1p family; potential Cdc28p substrate |

Table 3. Transcription factors that recognize TFBSs in Motifs 1 - 5.

The recognized consensus sequence, relative position and bound strand is indicated. For each TF the protein information deposited in SGD and Yeastract is provided.

## 4. Discussion

This work was initiated as an attempt to understand and define the promoter structure of the 10 *lip* genes from the ligninolytic basidiomycete *P. chrysosporium*, assuming that the members of this family are co-regulated and have a common code for this particular biological function. The first encouraging hint was the discovery of common TFBS sequences which suggested a coordinated response to the various processes involved in lignin biodegradation. Furthermore, the presence of a common organization might permit the identification of additional genes in the *P. chrysosporium* genome that participate in lignin degradation, on the basis that they received similar regulatory "inputs".

Multiple alignment of all *lip* promoters yielded short homologous sequences that included experimentally validated TFBSs in other eukaryotic organisms, including yeast. These results were very encouraging. Hoping to find that similar promoters would present comparable physiological responses, transcriptional levels of *lip* genes of the fungus grown in C- and N-limited cultures were examined. However, no clear correlation between genomic organization and transcript levels was observed under these conditions. Analysis of the 10 promoters using multiple programs and databases only showed scattered and ambiguous (or degenerate) TFBSs and no clear structural organization emerged. The use of MEME software represented a breakthrough, since it allowed finding sequences that share a common (but hidden) property in conserved positions, which do not correspond to *a priori* experimentally determined TFBSs (called *Ab Initio*). MEME detected the five relevant and statistically significant motifs presented in this work. This is consistent with the group of six *lip* genes with a highly conserved gene and protein structure, which had been previously reported by Stewart and Cullen (1999). This finding suggested that the subfamily I *lip* genes derived from several duplication events of an ancestral gene.

The next task consisted in determining if there is a common biological function associated to each motif. For this, the sequence of each motif was analyzed in the YEASTRACT database which identified yeast genes which also contained any of the five motifs within their regulatory sequences. Indeed, one or more genes were found for each motif which contained curated and experimentally validated TFBSs. How do these genes relate to the biological process of lignin biodegradation? In order to answer this question, each motif was analyzed.

Motif 1 included a single gene associated to the cellular response to oxidative stress. During secretion of enzymes involved in the ligninolytic process, such as LiPs and MnPs, oxidative

stress is a natural condition of *P. chrysosporium* and resistance to oxidative stress is probably an important function (Zacchi et al., 2000; Belinky et al., 2003; Jiang et al., 2005). To date there is no clear evidence in the literature on the mechanisms used by *P. chrysosporium* to tolerate the highly oxidative environment produced during lignin degradation. An orthologue of the yeast TRX2 gene, which encodes a cytoplasmic thioredoxin isoenzyme, could be involved in the protection of *P. chrysosporium* cells against oxidative and reductive stress. Motif 1 is also related to the secretion of vesicles, which is fully consistent with the manner in which these enzymes are carried into the extracellular medium.

Motif 2 is contained in several genes that do not share an obvious common function, although most of them are related to nitrogen metabolism, the Krebs cycle and ribosomal biogenesis. As is well known, LiPs are induced in response to low nitrogen and low carbon conditions, which suggests that the cell might be increasing protein synthesis, a necessary process for hyphal remodeling and growth.

Motif 3 is common to genes involved in cellular response to nitrogen and carbon metabolism, including gluconeogenesis. Some genes containing this motif are involved in the nitrogen cellular response to starvation and regulation of carbohydrate metabolic processes. There is a partial overlap of biological functions (though not of genes) with Motif 2, however, other interesting biological processes also seem to be involved: invasive growth in response to glucose limitation, which suggests remodeling of fungal cellular structures, such as cell wall assembly. It is known that during the ligninolytic process, *P. chrysosporium* apical tips of hyphae penetrate the wood through the tracheids and secrete ligninolytic enzymes. The yeast gene YDR477W | SNF1 contains motif 3 in its promoter and encodes an AMP-activated serine / threonine protein kinase, which is involved in signal transduction and found in a complex with proteins required for the transcription of glucose-repressed genes and involved in sporulation and peroxisome biogenesis. This gene might be related to stress tolerance regulation and gene expression under low carbon conditions, as would occur in secondary metabolism (ligninolysis), which is coupled to sporulation (structural remodeling of the fungus) and possibly, peroxisome biogenesis.

Motif 4 is present in the promoter of two yeast genes described as ion transporters which are of interest in relation to lignin biodegradation: genes YML121W | gtr1 and YML123C | PHO84 are involved in phosphate transport, which is essential for nucleic acids synthesis, and therefore also associated to cell cycle regulation, which in turn might be related to hyphal growth. YML121W | gtr1 encodes a cytoplasmic GTP binding protein and negative regulator of the Ran/Tc4 GTPase cycle; it is also a component of the GSE complex required for sorting of Gap1p and is involved in phosphate transport and telomeric silencing, similar to human Raga and Ragbir proteins. YML123C | PHO84 is a high-affinity inorganic phosphate (Pi) and low-affinity manganese transporter. The latter is relevant in the context ligninolysis since $Mn^{+2}$ has a regulatory role in the formation of LiPs (Rabinovich et al., 2004). Transport of this ion is important for the expression and activity of all kinds of ligninolytic enzymes from *P. chrysosporium*. Motif 4 is the most distal motif identified in the *lip* gene promoters. Due to its location on the promoter, it is tempting to speculate that this motif might be involved in DNA looping.

Motif 5 appears to be related to mitosis, cell cycle, chromosome segregation and stress response. The two yeast genes with motif 5 in their promoters and selected with the greatest stringency by YEASTRACT, YFR028C | CDC14 and YGR098C | ESP1, are required for the regulation of mitotic exit. This correlates well with active cell division that occurs in hyphae.

Other promoters which contain Motif 5, such as those from genes YDL079C | MRK1 (a glycogen synthase kinase 3 (GSK-3)) homolog and YOR181W|LAS17 are stress responsive genes. Finally, gene YGR274C | TAF1 (which encodes a TFIID subunit and is involved in promoter binding and G1 / S progression) and gene YOR140W | SFL1 are RNA polymerase II regulators. These functions appear to be complementary to those associated to the other motifs.

How do these regulatory elements coordinate fungal metabolism in natural environments? It is well known that filamentous fungi grow by apical extension and lateral branching to form mycelial colonies (Richards et al., 2010). Because of key characteristics of hyphae, filamentous fungi can efficiently colonize and exploit the substratum on which they grow, e.g. wood (Weber, 2002). Fungal cells within a single mycelium are known to autolyse to provide nutrients to ensure growth (Zacchi et al., 2000), involving processes related to the remodeling of the mycelium. In fungi, vacuoles are very versatile organelles involved in protein turnover, cellular homeostasis, membrane trafficking, signaling and nutrition (Veses et al., 2008), as well as progression through cell cycle checkpoints (Richards et al., 2010). Networks of spherical and tubular vacuoles have been found in a range of filamentous fungi, including the wood rotting plant pathogen *Phanerochaete velutina* (Richards et al., 2010). Under LiP producing conditions, hyphal cells undergo a major loss of cellular ultrastructure, similar to that observed under oxidative stress (Zacchi et al., 2000). Therefore LiPs may be enzymes that are induced under conditions of oxidative stress (Rabinovich et al., 2004) and degrade lignin in order to access further carbon sources (Zacchi et al., 2000).

Taken together, many of the genes shown to contain any of these motifs have in common that they regulate genes of relevance associated to the biological processes that occur during lignin biodegradation. They include stress, mycelia remodeling which involves changes in lipid and carbohydrate metabolism, and mitosis, that lead to organellar /ultrastructural reorganization and changes related to the shift to secondary metabolism. In an analogous manner, transcription factors that apparently recognize these motifs, also bind TFBSs of genes involved in stress response and mitosis, among others (See Table 3).

## 5. Final remarks and conclusion

This work proposes an ordered and step by step approach for the analysis of the putative structure of eukaryotic promoters. To test this strategy, the *lip* gene family from the ligninolytic fungus *P. chrysosporium* was studied. The resulting analysis uncovered an organization of TFBSs into structural motifs that is not evident using standard software. The MEME software, which searches for signals that are shared by a group of sequences, was instrumental to detect these hidden elements. Each of the discovered motifs contains several TFBSs. One transcription factor may bind to various sites and hence it is speculated that the TFBS pairs group into clusters, which may be bound by the same transcription factor. Clusters with TATA-related and CAAT-related pairs have been reported (Ma et al., 2004). Also, several TATA-box related triples have been described in the literature (Ma et al., 2004). Each motif found in our analysis may represent this clustering of TFBSs and therefore may correspond to the basic functional unit of a promoter. The functional promoter may then be an organized sequence of motifs, as diagrammed in Figures 5 and 6. A simple sentence can be envisioned as an analogy of this regulatory structure: a sentence containing an instruction in any language corresponds to a meaningful sequence of words. The promoter represents this sentence and each motif corresponds to one of the words. In turn, as each

word is composed by several syllables, each motif is built by combining several TFBSs. Just as syllables, which contain several letters, isolated TFBSs contain several nucleotides and may be present in more than one copy in a single word or appear in several different words within the same sentence, but often do not have functional meaning on their own.

In conclusion, this work proposes an ordered and step by step approach for the analysis of the putative structure of eukaryotic promoters. We devised a straightforward *in silico* strategy that permits the identification of promoter structure in a set of related eukaryotic genes. To test this strategy the *lip* gene family from the ligninolytic fungus *P. chrysosporium* was studied. The resulting analysis uncovered an organization of TFBSs into structural motifs (that are not evident using standard software) which are present in yeast genes and transcription factors involved in diverse processes related to the biological context in which ligninolysis is carried out. The structured motifs discovered in this study may represent a functional organization of regulatory sequences. A future challenge will be to test other gene families in order to determine if the proposed model is a general feature of eukaryotic systems.

## 6. Acknowledgments

We wish to thank Dr. Roberta Farrell, Waikato University, New Zealand, for revision of the manuscript.

## 7. References

Abdulrehman, D., Monteiro, P.T., Teixeira, M.C., Mira, N.P., Lourenço, A.B., dos Santos, S.C., Cabrito, T.R., Francisco, A.P., Madeira, S.C., Aires, R.S., Oliveira, A.L., Sá-Correia, I. & Freitas, A.T. (January 2011). YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface. *Nucleic Acids Research*, Vol. 39, No. (Database issue, Epub 2010 Oct 23), pp. D136–D140.

Bailey, T.L. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, USA.

Bailey, T.L. & Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*. Vol. 14, pp. 48-54.

Bailey, T.L., Bodén, M., Whitington, T., & Machanick, P. (2010). The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics,* Vol. 11 (April 2010), pp. 179.

Bailey, T.L., Williams, N. Misleh, C. & Li, W.W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acid Research,* Vol. 34, No. WebServer Issue, pp. W369-W373.

Belinky, P.A., Flikshtein, N., Lechenko, S., Gepstein, S., & Dosoretz, C.G. (2003). Reactive oxygen species and induction of lignin peroxidase in *Phanerochaete chrysosporium*. *Applied and Environmental Microbiology.* Vol. 69, No. 11 (November 2003), pp. 6500-6506.

Dashtban, M., Schraft, H. & Qin, W. (2009). Fungal conversion of lignocellulosic residues; opportunities & perspectives. *International Journal of Biological Sciences*, Vol. 5, No 6, pp. 578-595.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological),* Vol. 39, No. 1, pp. 1–38.

D'Souza, T.M., Merritt, C.S. & Reddy, C.A. (1999). Lignin-modifying enzymes of the white rot basidiomycete *Ganoderma lucidum*. *Applied and Environmental Microbiology,* Vol. 65, No. 12, pp. 5307-5313.

Farrell, R.L., Murtach, K.B., Tien, M, Mozuch, M.D., & Kirk, T.K. (1989). Physical and Enzymatic Properties of Lignin Peroxidase isoenzymes from *Phanerochaete chrysosporium*. *Enzyme Microbial Technology,* Vol. 11, No. 6, pp. 322-28.

Han, L., Garcia, H.G., Blumberg, S., Towles, K.B., Beausang, J.F. & Phillips, R. (2009). Concentration and length dependence of DNA looping in transcriptional regulation *Plos One,* Vol. 4, No 5., pp. e5621.

Haygood, R., Babbitt, C.C., Fedrigo, O. & Wray, G.A. (2010). Contrasts between adaptative coding and noncoding changes during human evolution. *Proceedings of the National Academy of Sciences.* Vol. 107, No. 17 (May 2010), pp. 7853-7857.

Hein, J. (1990) Unified approach to alignment and phylogenies. *Methods in Enzymology*, Vol. 183, pp. 626-645, Academic Press, Inc. San Diego, California, USA.

Howard, M.L. & Davidson, E.H. (2004). Cis-Regulatory control circuits in development. *Developmental Biology*, Vol. 271, No. 1, pp. 109-118.

Jiang, Q., Yan, Y.-H., Hu, G.-K. & Zhang, Y.-Z. (2005). Molecular cloning and characterization of a peroxiredoxin from *Phanerochaete chrysosporium. Cellular and Molecular Biology Letters*, Vol. 10, No. 4, pp. 659 – 668.

Jiang, X., Zeng, G., Huang, D., Chen, Y., Liu, F., Huang, G., Li, J., Xi, B. & Liu, H. (2006) Remediation of pentachlorophenol-contaminated soil by composting with immobilized *Phanerochaete chrysosporium*. *World Journal of Microbiology and Biotechnology*, Vol. 22, No 9, pp. 909-913.

Kirk, T.K. & Farrell R.L. (1987). Enzymatic "combustion": the microbial degradation of lignin. *Annual Review of Microbiology,* Vol. 41, pp. 465-505.

Kirk, T.K., Tien, M., Kersten, P., Kalyanaraman, B., Hammel, K., & Farrell, R.L. (1990). Lignin Peroxidase from Fungi *Phanerochaete chrysosporium*. *Methods of Enzymology*, Vol. 188, p. 159-171, Academic Press, Inc. San Diego, California, USA.

Lobos, S., Larraín, J., Salas, L., Cullen, D. & Vicuña, R. (1994). Isoenzymes of manganese-dependent peroxidase and laccase produced by the lignin degrading basidiomycete *Ceriporiopsis subvermispora. Microbiology,* Vol. 140, pp. 2691–2698.

Ma, X.-T., Qian, M.-P. & Tang H.-X. (2004). Predicting Polymerase II Core Promoters by Cooperating Transcription Factor Binding Sites in Eukaryotic Genes. *Acta Biochimica et Biophysica Sinica,* Vol. 36, No. 4, pp. 250–258.

Mariño-Ramírez, L., Spouge, J.L., Kanga, G.C. & Landsman, D. (2004). Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Research*, Vol. 32, No. 3, pp. 949-958.

Martínez, D., Larrondo, L.F., Sollewijn, G., Huang, K., Helfenbein, K.G., Ramaiya, P., Detter, J.C., Larimer, F., Coutinho, P.M., Henrissat, B., Berka, R., Cullen, D., & Rokhsar, D.

(2004). Genome sequence of the lignocelluloses degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nature Biotechnology*, Vol. 22, pp. 695-700.

Martínez, D., Challacombe, J., Morgenstern, I., Hibbett, D., Schmoll, M., Kubicek, C.P., Ferreira, P., Ruiz-Duenase, F.J., Martinez, A.T., Kersten, P., Hammel, K., Wymelenbergg, A.V., Gaskell, J., Lindquist, E., Sabat, G., Splinter, S., BonDuranti, Larrondo, L.F., Canessa, P., Vicuña, R., Yadav, J., Doddapaneni, H., Subramanian, V., Pisabarro, A.G., Lavín, J.L., Oguiza, J.A., Master, E., Henrissat, B., Coutinho, P.M., Harris, P., Magnuson, J.K., Baker, S.E., Bruno, K., Kenealy, W., Hoegger, P.J., Kües, U., Ramaiya, P., Lucas, S., Salamov, A., Shapiro, H., Tu, H., Chee, C.L., Misra, M., Xie, G., Teter, S., Yaver, D., James, T., Mokrejs, M., Pospisek, I.V. Grigoriev, T. Brettin, D. Rokhsar, Berka, R. & Cullen, D. (2009). Genome, transcriptome, and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion. *Proceedings of the National Academy of Sciences*, Vol. 106, No. 6, pp. 1954–1959.

Matthews, K.S. (1992). DNA looping. *Microbiology Review,* Vol. 56, No. 1, pp. 123-136.

Monteiro, P.T., Mendes, N.D., Teixeira, M.C., d'Orey, S., Tenreiro, S., Mira, N.P., Pais, H., Francisc, A.P., Carvalho, A.M., Lourenço, A.B., Sá-Correia, I., Oliveira, A.L. &Freitas, A.T. (2008). YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in *Saccharomyces cerevisiae. Nucleic Acids Research*, Vol. 36, No. 36 Database issue (January 2008), pp. D132-D136.

Patikoglou, G.A., Kim, J.L., Sun, L., Yang, S.H., Kodadek, T. & Burley, S.K. (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes and Development,* Vol. 13, (Dec 1999), No. 24, pp. 3217-3230.

Rabinovich, M.L., Bolobova, A.V. & Vasil'chenko L.G. (2004). Fungal decomposition of natural aromatic structures and xenobiotics: a review. *Applied Biochemistry and Microbiology*, Vol. 40, No. 1, pp. 1-17.

Rajakumar, S., Gaskell, J., Cullen, D, Lobos, S, Karahanian, E. & Vicuña, R. (1996). Lip-Like genes in *Phanerochaete sordida* and *Ceriporiopsis subvermispora*, white rot fungi with no detectable lignin peroxidase activity. *Applied and Environmental Microbiology*, Vol. 62, No. 7 (July 1996), pp. 2660–2663.

Richards, A., Veses, V. & Gow, N.A.R. (2010). Vacuole dynamics in fungi. *Fungal Biology Reviews*, Vol. 24, pp. 93-105.

Rüttimann-Johnson, C., Cullen, D. & Lamar, R. (1994). Manganese peroxidases of the white rot fungus *Phanerochaete sordida. Applied and Environmental Microbiology*, Vol. 60, pp. 599–605.

Sánchez, C. (2009). Lignocellulosic residues: Biodegradation and bioconversion by fungi. *Biotechnology Advances*, Vol. 27, No. 2 (March-April 2009), pp. 185-194.

Sato, S., Liu, F., Koc, H. & Tien, M. (2007). Expression analysis of extracellular proteins from *Phanerochaete chrysosporium* grown on different liquid and solid substrates. *Microbiology,* Vol. 153, No. Pt 9 (September 2007), pp. 3023-3033.

Schleif, R. (1992). DNA looping. *Annual Review of Biochemistry*, Vol. 61, pp. 199-223.

Shi, J., Chinn, M.S. & Sharma-Shivappa, R.R. (2008). Microbial pretreatment of cotton stalks by solid state cultivation of *Phanerochaete chrysosporium. Bioresources Technology,* Vol. 99, No. 14 (September 2008), pp 6556-6564.

Shi, J., Sharma-Shivappa, R.R. & Chinn, M.S. (2009). Microbial pretreatment of cotton stalks by submerged cultivation of *Phanerochaete chrysosporium*. *Bioresources Technology,* Vol. 100, No. 19 (October 2009), pp. 4388-4395.

Singh, D. & Chen, S. (2008). The white-rot fungus *Phanerochaete chrysosporium*: conditions for the production of lignin-degrading enzymes. *Applied Microbiology and Biotechnology*, Vol. 81, No. 3 (December 2008), pp. 399-417.

Singh, P., Sulaiman, O., Hashim, R., Rupani, P.F. & Peng, L.C. (2010). Biopulping of lignocellulosic material using different fungal species. *Reviews in Environmental Science and Biotechnology*, Vol. 9, pp. 141-151.

Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, Vol. 147, No. 1, pp. 195-197.

Stepanova, M., Tiazhelova, T., Skoblov, M. & Baranova, A. (2005). A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas *Bioinformatics,* Vol. *21, N*o. 9 (May 2005), pp. 1789–1796.

Stewart, P. & Cullen, D. (1999). Organization and Differential Regulation of a cluster of lignin Peroxidase genes from *Phanerochaete chrysosporium. Journal of Bacteriology,* Vol. 181, No. 11 (June 1999), pp. 3427-3432.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix. *Nucleic Acids Research*, Vol. 22, No. 22 (November 1994), pp. 4673-4680.

Veses, V., Richards, A. & Gow, N.A.R. (2008). Vacuoles and fungal biology. *Current Opinion in Microbiology*, Vol. 11, pp. 503-510.

Weber, R.W.S. (2002). Vacuoles and the fungal lifestyle. *Micologist*, Vol. 16. Part 1, pp. 10-20.

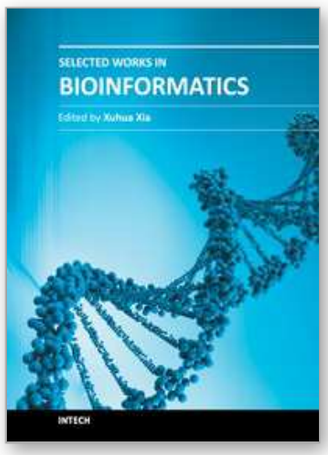
Wei, W. & Yu. X.-D. (2007). Comparative analysis of regulatory motif discovery tools for Transcription Factor Binding Sites. *Genomics Proteomics Bioinformatics,* Vol. 5, No. 2 (May 2007*),* pp. 131-142.

Wittkopp, P.J. (2006). Evolution of *cis*-regulatory sequence and function in Diptera. *Heredity,* Vol. 97, No. 3 (July 2007), pp. 139-147.

Wu, J.M. & Zhang, Y.Z. (2010). Gene Expression in Secondary Metabolism and Metabolic Switching Phase of *Phanerochaete chrysosporium*. *Applied and Biochemical Biotechnology*, Vol. 162, No. 7 (November 2010), pp. 1961-1977.

Wymelenberg, A.V., Gaskell, J., Mozuch, M., Kersten, P., Sabat, G., Martinez, D. & Cullen, D. (2009). Transcriptome and Secretome Analyses of *Phanerochaete chrysosporium* Reveal Complex Patterns of Gene Expression. *Applied and Environmental Microbiology*, Vol. 75, No. 12 (June 2009), pp. 4058–4068.

Zacchi, L. Morris, I. & Harvey, P.J. (2000). Disordered ultrastructure in lignin-peroxidase-secreting hyphae of the white-rot fungus. *Phanerochaete chrysosporium. Microbiology*, Vol.146, pp. 759-765.




**Selected Works in Bioinformatics**

Edited by Dr. Xuhua Xia

ISBN 978-953-307-281-4
Hard cover, 176 pages
**Publisher** InTech
**Published online** 19, October, 2011
**Published in print edition** October, 2011

This book consists of nine chapters covering a variety of bioinformatics subjects, ranging from database resources for protein allergens, unravelling genetic determinants of complex disorders, characterization and prediction of regulatory motifs, computational methods for identifying the best classifiers and key disease genes in large-scale transcriptomic and proteomic experiments, functional characterization of inherently unfolded proteins/regions, protein interaction networks and flexible protein-protein docking. The computational algorithms are in general presented in a way that is accessible to advanced undergraduate students, graduate students and researchers in molecular biology and genetics. The book should also serve as stepping stones for mathematicians, biostatisticians, and computational scientists to cross their academic boundaries into the dynamic and ever-expanding field of bioinformatics.



INTECH
open science | open minds

**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen