

# Predicting *cis*-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups

Michiel Wels<sup>1,2,3,\*</sup>, Christof Francke<sup>1,2</sup>, Robert Kerkhoven<sup>2</sup>, Michiel Kleerebezem<sup>1,3</sup>  
and Roland J. Siezen<sup>1,2,3</sup>

<sup>1</sup>Wageningen Centre for Food Sciences, PO Box 557, 6700 AN Wageningen, The Netherlands, <sup>2</sup>Radboud University, Centre for Molecular and Biomolecular Informatics, PO Box 9010, 6500 GL Nijmegen, The Netherlands and <sup>3</sup>NIZO food research, PO Box 20, 6710 BA Ede, The Netherlands

Received February 10, 2006; Revised March 1, 2006; Accepted March 15, 2006

## ABSTRACT

*Cis*-acting elements in *Lactobacillus plantarum* were predicted by comparative analysis of the upstream regions of conserved genes and predicted transcriptional units (TUs) in different bacterial genomes. TUs were predicted for two species sets, with different evolutionary distances to *L. plantarum*. TUs were designated 'cluster of orthologous transcriptional units' (COT) when >50% of the genes were orthologous in different species. Conserved DNA sequences were detected in the upstream regions of different COTs. Subsequently, conserved motifs were used to scan upstream regions of all TUs. This method revealed 18 regulatory motifs only present in lactic acid bacteria (LAB). The 18 LAB-specific candidate regulatory motifs included 13 that were not described previously. These LAB-specific different motifs were found in front of genes encoding functions varying from cold shock proteins to RNA and DNA polymerases, and many unknown functions. The best-described LAB-specific motif found was the CopR-binding site, regulating expression of copper transport ATPases. Finally, all detected motifs were used to predict co-regulated TUs (regulons) for *L. plantarum*, and transcriptome profiling data were analyzed to provide regulon prediction validation. It is demonstrated that phylogenetic footprinting using different species sets can identify and distinguish between general regulatory motifs and LAB-specific regulatory motifs.

## INTRODUCTION

Many microorganisms are able to survive in environments where conditions change rapidly. Appropriate and fine-tuned environmental responses require gene regulatory networks that are efficient, flexible, robust and contain internal controls and feedback mechanisms, to avoid overreaction to certain stimuli.

The comprehensive interpretation of gene expression data can be greatly enhanced by an understanding of regulatory networks. Such understanding could elucidate regulatory processes underlying specific *in situ* behavior, e.g. during gastro-intestinal tract residence or during food fermentation processes, providing targets for optimizing culture performance and improving strain robustness. By pinpointing possible bottlenecks in the regulatory network, it may be possible to modulate a whole pathway by knocking out or over-expressing only a single regulatory gene. Insight in gene regulatory networks can be derived from experimental post-genomics approaches such as transcriptome profiling, which can reveal co-regulated genes (regulons) and regulatory networks that are present in a specific microorganism (1,2).

Another, potentially more generic way to obtain insight in the regulatory network of one or more organisms is by *in silico* detection of (conserved) *cis*-acting elements, representing the DNA-binding sites for regulatory proteins (*trans*-acting elements). Using this approach, potential regulons can be identified on basis of shared *cis*-acting elements preceding the co-regulated genes. The identification of regulons can enhance the insight in gene-function relations and elucidate mechanisms underlying adaptation to changing environmental conditions. In various *in silico* studies, *cis*-acting elements have been predicted in bacterial genomes (3,4). The upstream regions of a group of genes predicted to have the same *cis*-acting element (for instance on basis of micro-array data or

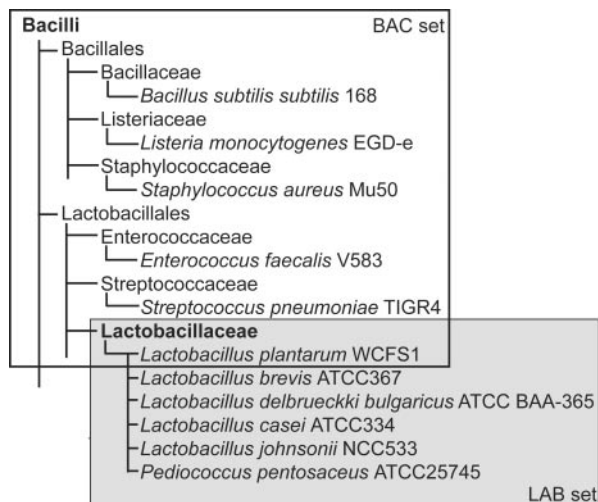
\*To whom correspondence should be addressed at CMBI, Radboud University, PO Box 9010 6500 GL Nijmegen, The Netherlands. Tel: +31 024 3653398; Fax: +31 024 3652977; Email: M.Wels@cmbi.ru.nl

sequence conservation) can be analyzed using pattern recognition tools such as Gibbs sampling (5) or expectation maximization (EM) (6). Subsequently, conserved motifs in the upstream regions can be used to scan the genome(s) of interest in order to predict regulons. Alternatively, comparative genomics can be used for the identification of genes with conserved *cis*-acting elements in different species, including the identification of regulons within a single species. The underlying assumption is that orthologous genes in different organisms are regulated in a similar manner (7). Orthologous genes can be identified in different species by using orthology prediction methods such as COG (8). This method, known as phylogenetic footprinting, was successfully applied for detection of *cis*-acting elements in different (sets of) species (9–12). Moreover, recent evidence shows that *cis*-acting elements predicted with a small species set can be verified by searching for these elements in the upstream regions of orthologous genes in other genomes, that were not incorporated in the initial set (11).

Since initial phylogenetic footprinting is often performed with small species sets, the selection of species is of utmost importance. Species that are phylogenetically too distantly related will provide problems in the orthology prediction, and will only allow detection of generally well-described motifs upstream of highly conserved genes. In contrast, comparing too closely related species tends to generate a higher frequency of false-positive motifs due to high-level conservation of intergenic region sequences, which hampers detection of candidate *cis*-acting elements.

Lactic acid bacteria (LAB) are industrially important microorganisms that can be found in several starter cultures for food and feed fermentations as well as in the human gastrointestinal tract (13). *Lactobacillus plantarum* is an exemplary LAB that is encountered in and, therefore, able to adapt to many different environmental niches. The genome of *L.plantarum* is considered to be among the largest of *Lactobacilli* (14) and it is postulated to encode a relatively large number of regulatory proteins in comparison to other LAB (15), including other *Lactobacillaceae* like *Lactobacillus johnsonii* (16).

In this work, an in-depth phylogenetic footprinting analysis was performed on the complete genome sequence of *L.plantarum* WCFS1 (15) to identify regulatory networks. Availability of (partial) genome sequences of many closely related species, as well as more distant species, allowed the determination of the effect of two different species sets on phylogenetic footprinting results. Both sets consisted of six different species, for which the average evolutionary distance to *L.plantarum* differed (Figure 1). In the first set (BAC set), species were chosen from different families of the class of *Bacilli*. Next to *L.plantarum* (*Lactobacillaceae*), one species was selected from all families of which at least one completely sequenced genome was available (*Bacillaceae*, *Enterococcaceae*, *Listeriaceae*, *Staphylococcaceae* and *Streptococcaceae*). The other species set (LAB set) was a selection of genomes that only represent the family of *Lactobacillaceae* (including *Pediococcus pentosaceus*, which is also considered to be a member of the family of *Lactobacillaceae*). By using these different species sets, we identified and distinguished between regulatory motifs conserved among *Bacilli* and/or *Lactobacillaceae*. All motifs



**Figure 1.** Phylogenetic relation of species gathered from the TaxBrowser at NCBI. Relations are based on 16S rRNA sequence. Different species sets (BAC set and LAB set) were chosen on basis of the phylogenetic distance to *L.plantarum*. All members of the LAB set are more closely related to *L.plantarum* than to members of the BAC set.

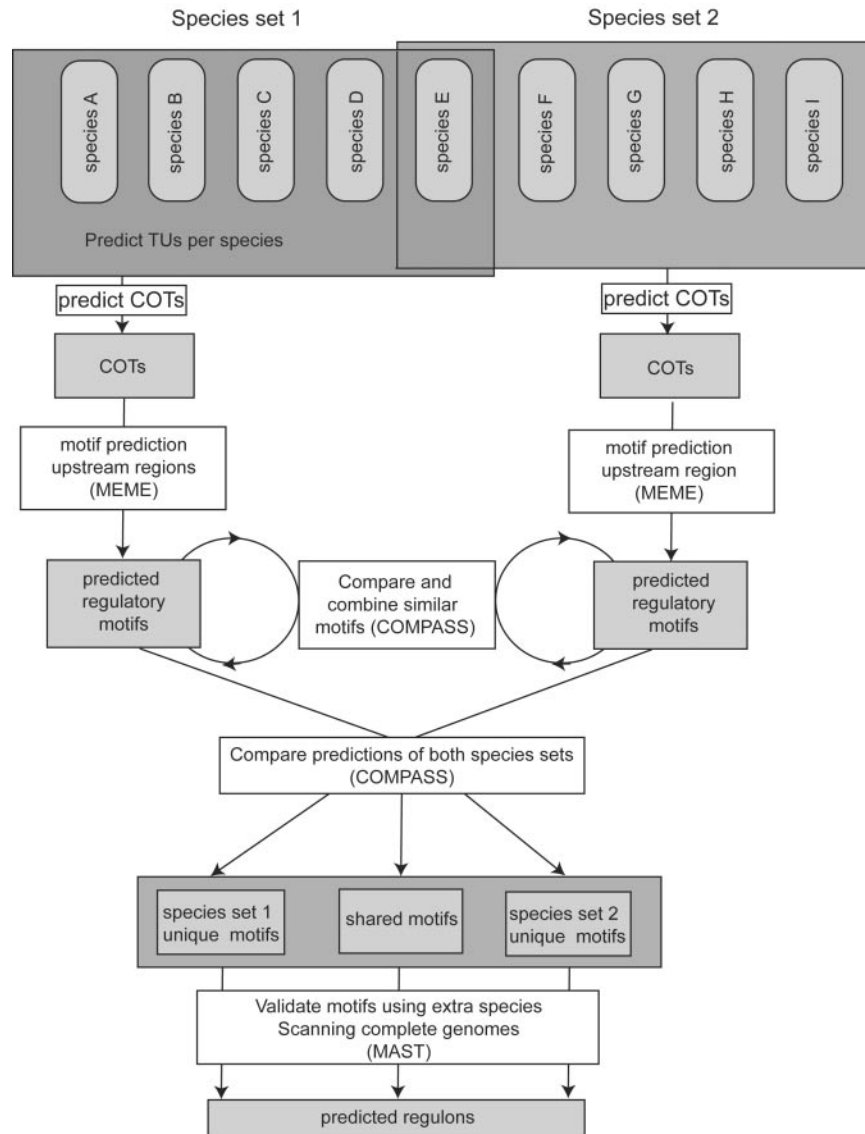
detected were used to predict regulons encoded by the *L.plantarum* genome. The availability of a growing set of microarray data of *L.plantarum* in our laboratory enabled calculation of expression correlations for selected genes and allowed validation of several regulon predictions provided by phylogenetic footprinting.

## MATERIALS AND METHODS

A schematic representation of the phylogenetic footprinting procedures employed is depicted in Figure 2.

### Species selection

The *Bacilli* set (BAC set) consisted of the following organisms: *Streptococcus pneumoniae* TIGR4, *Bacillus subtilis* 168, *Staphylococcus aureus* Mu50, *L.plantarum* WCFS1, *Enterococcus faecalis* V583 and *Listeria monocytogenes* EGD-e. The genomic information (genome sequence and gene predictions) for these organisms was taken from public databases (Genbank, <http://ncbi.nlm.nih.gov>). The genomes of the *Lactobacillaceae* set (LAB set), consisted of the organisms *L.plantarum* WCFS1, *L.johnsonii* NCC533, *Lactobacillus brevis* ATCC367, *Lactobacillus delbrueckii* ssp. *bulgaricus* ATCC BAA-365, *Lactobacillus casei* ATCC334 and *P.pentosaceus* ATCC25745. If not available in public databases (Genbank <http://www.ncbi.nlm.nih.gov/>), the data were taken from the ERGO bioinformatics suite (17). At the time of our analysis, the latter four genomes were from unfinished sequencing projects from the Joint Genome Institute [http://genome.jgi-psf.org/mic\\_cur1.html](http://genome.jgi-psf.org/mic_cur1.html) with the genomic information being retrieved from several contigs. Further comparisons were made with all publicly available completed genomes and the incomplete genomes of *Enterococcus faecium* DO, *Lactobacillus gasseri* ATCC-33323, *Oenococcus oeni* PSU-1 and *Leuconostoc mesenteroides* ATCC-8293, all available in the Genbank database.



**Figure 2.** Schematic representation of the motif prediction procedure employed in this study. For each species TUs were predicted and used for a COT prediction. The upstream regions of the COTs were analyzed using MEME. Following MEME analysis, predicted regulatory motifs were compared using COMPASS. The upstream regions containing significantly similar motifs were re-analyzed by MEME. This procedure was iterated until all identified motifs could be considered unique. The unique motifs of both sets were compared and on basis of this comparison, the motifs were divided into three different classes. All motifs were validated using MAST against other genomic sequences. Finally, regulons were predicted by scanning the genome with the identified motifs.

### Transcriptional unit (TU) prediction

TU predictions have been performed before for several of these genomes (18), but not all. Therefore, TU predictions were performed for all species in the two different species sets. TU prediction was based on three genome context parameters: genes were considered to be present in the same TU if (i) adjacent genes were positioned on the same coding strand, (ii) adjacent genes had an intergenic region <100 bp, and (iii) no TransTerm (19) predicted (Rho-independent) termination signal was present between adjacent genes.

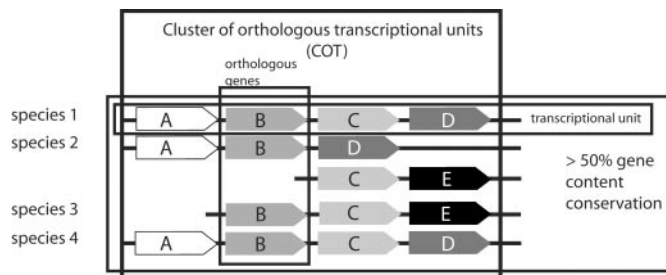
### Orthology prediction

Orthologous genes were predicted with a sensitive search using the Smith-Waterman algorithm (20) against the COG

(Clusters of Orthologous Groups) database (21). The search was performed with the following parameter settings for the SW algorithm; gap penalty,  $-1$ ; gap opening,  $-11$ ; scoring matrix, *blosum62*. When the best hits of the protein against the COG database were all part of the same COG, this COG was assigned to the protein. In some other situations, multiple COGs were assigned to one protein; by allowing different COGs to be assigned to different parts of the protein, orthology prediction of fusion proteins was achieved.

### COTs prediction

TU and protein-orthology predictions were combined to predict conserved orthologous transcriptional units (COTs) (Figure 3). If >50% of the genes of the smallest of the two



**Figure 3.** Prediction of the COTs. The TUs of the different species were compared. If >50% of the gene content of the smallest TU was shared, the TUs were considered to be orthologous and combined into one cluster. Gene order was allowed to vary in the analysis.

compared TUs were orthologous to genes in the other TU, these TUs were considered orthologous. At least three out of the six species from a set had to be represented by at least one TU to be considered a COT. The COTs also contained TUs with indirect relations. For example, in Figure 3, the first TU of species 2 and the TU in species 3 share <50% orthology. Still they are classified to the same COT, since the TU in species 1 and species 4 share >50% of orthology with both the first TU in species 2 as well as the TU in species 3. In addition, in one COT multiple TUs can be present per species; the first TU of species 2 shares >50% of orthologous genes with the TU in species 1 and 4, while the second TU shares >50% of orthologous genes with the TU in species 3. Gene order variation within a TU was allowed.

### Upstream region comparison

The upstream DNA regions of all TUs in one COT were compared using MEME (6), an EM based algorithm. Upstream regions of 300 bp in length were used to detect *cis*-acting elements, unless the intergenic region was smaller, in which case the total intergenic region was used as input, with a minimum of 50 bp. Twenty base pairs of the coding sequence of the first gene of each TU was included in the analysis, to correct for errors in the predictions of gene starts. Since the first genes of TUs within a COT are often orthologous and sequence conservation within these orthologs is better conserved than within the upstream sequence, the region within the gene was maximized to 20 bp. In this way MEME only predicts one motif as a result of functional domain conservation within genes. Since the LAB set contained genomes that were not fully sequenced, the upstream regions of three different species had to be present to be analyzed by MEME. MEME has two major advantages in comparison to other pattern recognition programs, that make it suitable for *ab initio* prediction: (i) MEME can predict motifs of different lengths, depending on a selected length range. If the length of a *cis*-acting element is not known, MEME will generate a motif with the correct size, and (ii) MEME can search for multiple motifs in a given set of sequences. If a COT has several conserved *cis*-acting elements (and is predicted to be regulated by several transcription factors), MEME will find them all.

The following MEME settings were used: search for motifs with a length ranging from 10 to 30 bp, maximally five motifs per COT and only on the COT coding strand. The statistical

parameter ZOOPS (Zero or One Occurrence Per Sequences) was given as an input, since it is not clear if all TUs of one COT had the same *cis*-acting element (owing to, for instance, errors in the COT prediction). A disadvantage of MEME is that it does not allow gaps within motifs, so that motifs that have a variable spacing region in between a direct or tandem repeat will be found as two separate *cis*-acting elements.

### Comparison of the predicted *cis*-elements

The multiple alignment comparison program COMPASS (22) was used to analyze and compare the predicted motifs from both species sets. COMPASS was originally written for comparing protein profiles but a simple change of the comparison matrix (from *blosum62* to a standard DNA matrix) makes it suitable for comparing DNA alignments. A multiple alignment was built for each predicted motif from the original MEME output. An all-against-all comparison was performed for the predicted motifs. In this way, over-represented regulatory motifs (predicted to be present in the upstream regions of several COTs in one species set) as well as the motifs predicted by both species sets could be detected. To reduce the number of false positives, only motifs with an *E*-value lower than 0.1 were selected. If similar motifs were found for more than one COT in one species set, the upstream regions were combined and a new MEME analysis was performed to refine motif predictions.

### Identification of relevant motifs

The predictions were compared with known *cis*-acting elements from literature. For each COT with a predicted conserved upstream motif, the DBTBS (23) was searched for a documented *cis*-element for the genes in the *B.subtilis* TU. If a *cis*-element was found, it was compared with the predicted motif. As the DBTBS only contains information on known *B.subtilis* regulatory elements, motifs that are either not conserved or not described in *B.subtilis* will be missed. Therefore, if the motif was not found in *B.subtilis* or there was no match in the DBTBS, the Pubmed database (<http://www.ncbi.nlm.nih.gov/entrez/>) was searched using the gene names of all genes of the different TUs. All COT-related abstracts were retrieved and manually scanned for experimental regulatory element information.

### Regulon prediction for *L.plantarum*

The identified motifs were searched in the genomes of different species using MAST (24). The MAST output was empirically filtered on basis of *P*-value, which represents the significance of a hit at a specific position of one of the sequences. As a cut-off, all hits with a *P*-value <  $10 \times 10^{-9}$  were considered positive. Hits with a *P*-value above  $10 \times 10^{-9}$  but below  $10 \times 10^{-5}$  were considered false-positive if the *P*-value of the hit divided by the *P*-value of the worst positive hit was higher than 100. In addition to the *P*-value cut-off, TUs were only considered valid members of the regulon if at least two members of one COT were found to have a significant hit with the motif.

### Expression correlation calculation

Data were obtained from 37 independent microarray experiments of *L.plantarum* WCFS1 using Agilent oligo-based

arrays. The tested conditions in the experiments differed from stress conditions to knockout or overexpression of metabolic genes (D. Molenaar, unpublished data). The  $^2\log(\text{cy}3/\text{cy}5)$  (M-value) was used to determine the correlation of expression between all possible gene pairs within the genome of *L.plantarum*. To reduce noise, only genes with an M-value variance  $>0.35$  were used. After applying this filter, 1998 out of 3024 predicted *L.plantarum* genes were suitable to use for TU and regulon validations. For each gene pair, the uncentered Pearson correlation was calculated (25). Correlations for subsets of genes (e.g. all correlations for genes belonging to one TU) were compared with all correlations.

## RESULTS

### Conserved orthologous TU prediction

TU predictions were performed on all genomes of both species sets (Table 1). This led to the prediction of 9618 TUs for the BAC set, with a mean of 1.86 genes per TU and 6464 TUs,

**Table 1.** Characteristics for each species set

	BAC	LAB
Number of species	6	6
Number of genes	17 922	11 436
Genes/species (mean)	2987	1906
Number of Tus	9618	6464
Genes/TU (mean)	1.86	1.77
TUs/species (mean)	1603	1077
Number of COTs	775	527
TUs/COT (mean) <sup>a</sup>	5.2	5.0
Number of unique motifs ( $E$ -value $< 0.1$ )	195	195

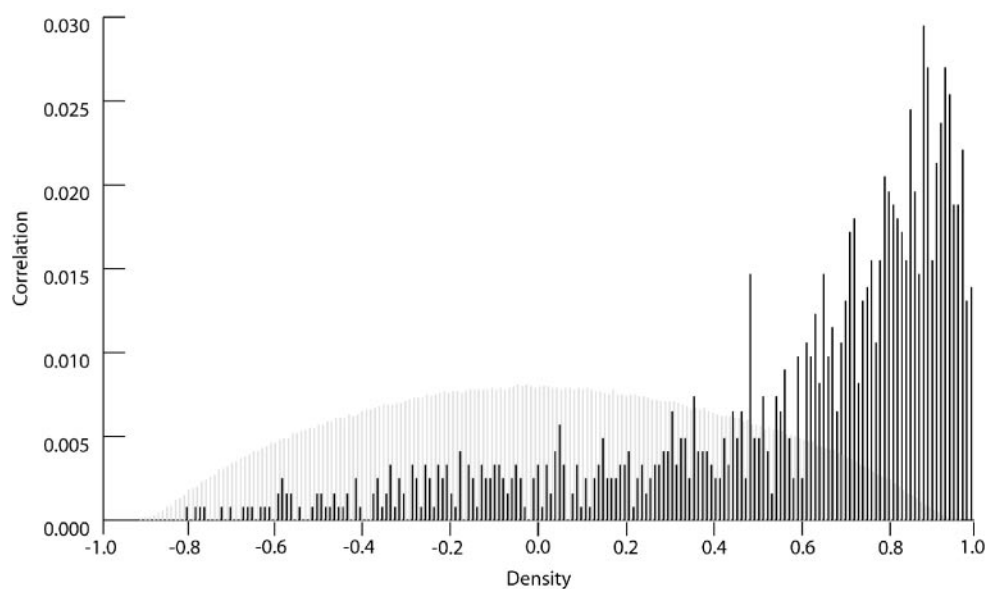
<sup>a</sup>Since a TU could be present in several COTs the mean number of TUs per COT is not equivalent to the total number of TUs divided by the total number of COTs.

with 1.77 genes per TU for the LAB set. The number of TUs per set differs due to large differences in genome size (the genomes in the BAC set were  $\sim 30\%$  larger than the genomes in the LAB set). The TU prediction in *L.plantarum* was validated using the gene expression data. Expression of pairs of genes within TUs was compared with gene pairs that were predicted not to be present in the same TU. Figure 4 shows the distribution of expression correlations of predicted TU gene pairs compared with all gene pairs. This analysis supports the TU prediction, since the expression of genes within a TU is found to correlate considerably better than the expression of random genes. Of all correlations within TUs 70% is  $>0.50$ , while only 15% of the complete dataset correlates  $>0.50$ .

Clusters consisting of TUs with  $>50\%$  of orthologous genes (COT) were formed by combining the TU and COG predictions (Figure 3): 775 different COTs were predicted for the BAC set, and 527 for the LAB set. Since 4 out of 6 genome sequences in the LAB set were incomplete, it is likely that less COTs will be predicted for this set. The mean number of TUs per COT is 5.2 for the BAC set and 5.0 for the LAB set.

### Comparative motif prediction by phylogenetic footprinting

The upstream regions of the COTs were used to predict conserved *cis*-acting elements. To increase reliability a COT was only analyzed if they were found in at least three species of the individual species sets, and if the corresponding upstream intergenic regions were at least 50 bp in length. Taking these selection criteria, *cis*-acting elements were searched in the upstream regions of 424 COTs of the LAB set and 652 COTs of the BAC set, using MEME with a positive-hit cut-off  $E$ -value of 0.1. Positively selected upstream regions of COTs were combined and re-analyzed by COMPASS, resulting in one general motif prediction for these COTs. Overall, this



**Figure 4.** Distribution of expression correlations for gene pairs of *L.plantarum*. The correlation of gene pairs within a TU (black), shows a clearly different distribution than the correlation of all gene pairs (gray).

resulted in the identification of 390 different motifs, equally divided over both species sets. These motifs were subsequently used to scan the upstream regions of all TUs in both species sets, using MAST. A simple filter was applied to distinguish shared and unique motifs from the complete set of motifs. A motif was considered present in a species, if it was found in at least one upstream region. If the motif is present in four out of six species of a species set, the motif was considered to be present in this set. If the motif could not be found in any species of a set or was found only once (generally *L.plantarum*, see below) the motif was considered absent in the species set. On basis of these criteria, motifs were classified as being unique (present in only one species set) or shared (present in both species sets), while motifs with a less clear distribution among the two species sets were not further characterized. Of the 195 motifs detected in the LAB species set, only 41 were also found in at least four BAC set species and therefore classified as shared. In analogy, 61 of the 195 motifs detected in the BAC species set were classified as shared between the two species sets. COMPASS-based analysis was used to compare these shared motifs: the highest scoring shared motifs and their predicted regulons in *L.plantarum* are listed in Table 2. In addition, eighteen motifs classified as unique for the *Lactobacillales* (LAB set; Table 3) and three motifs unique for the BAC set were identified. This finding supports the hypothesis that regulatory networks display better conservation in more closely related species. The full set of predicted motifs can be found at [https://bamics3.cmbi.ru.nl/cis\\_prediction](https://bamics3.cmbi.ru.nl/cis_prediction). While many of the predicted motifs correspond to experimentally validated regulatory elements (see examples below), the large majority of our predicted motifs are novel, and provide a wealth of targets and support for experimental data verification. Some of the shared and specific motifs will be discussed in further detail below.

### Recovery of well described regulatory motifs

**T-boxes.** COMPASS analysis showed that a large number of predicted motifs from both species sets are very similar. These motifs were identified as T-boxes. T-boxes are *cis*-acting elements affecting the translation of genes involved in aminoacyl-tRNA ligation and amino acid biosynthesis in many Gram-positive bacteria. After transcription, the T-box of the mRNA can fold into two different tertiary structures, depending on an interaction with the unloaded tRNA. If there are many unloaded tRNA molecules, they will bind to the codon part of the T-box which then folds into a structure that promotes translation. If many tRNAs are loaded, they have no interaction with the T-box, and the T-box will fold into the tertiary structure that inhibits translation of the mRNA [for a review see (26)].

Many of the predicted T-boxes were identified in front of COTs containing at least one gene encoding an aminoacyl-tRNA ligase. Other T-boxes appeared to be present upstream of COTs involved in amino acid biosynthesis and/or amino acid transport. It has been shown that a T-box is recognized by a specific tRNA, e.g. if a gene encodes a methionine aminoacyltransferase, the T-box upstream of this gene will have a specific interaction with the methionine tRNA. Moreover, amino acid specificity of the aminoacyltransferase

encoded appears to be conserved in the tRNA that binds to the corresponding upstream T-box sequence [for a review see (26)]. Since tRNA and corresponding amino acid specificity of the different T-boxes is determined by only three nucleotides within the T-box sequence (base pairing with the anticodon of the corresponding tRNA), this specificity was not immediately reflected in the identified *cis*-elements.






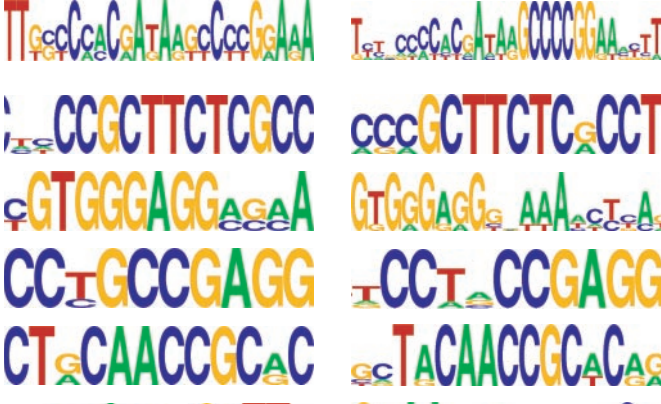




**CIRCE.** Another highly conserved and well-known regulatory motif is the CIRCE element, which is recognized by HrcA that regulates class I stress protein-encoding genes. Active CIRCE sites have been found upstream of stress response genes such as *groES/EL*, *dnaK* and *grpE* (27,28). When HrcA binds to CIRCE, it inhibits the expression of these stress genes. In this study, the CIRCE motif was confirmed to be present upstream of different COTs that contained stress response genes. In addition, the CIRCE element was detected upstream of the *hrcA* gene itself in all used species of both species sets, which is in good agreement with the observed *hrcA* auto-regulation in *B.subtilis* (27) and *Lactococcus lactis* MG1363 (29). Another COT, encoding a metal-dependent membrane-bound protease, often found divergently orientated in front of the TU containing *groES/EL*, is also predicted to be regulated by HrcA. Since the CIRCE element is a palindromic sequence, HrcA can bind on both the plus and minus strands. This suggests that a single CIRCE element could regulate the expression of divergent TUs. However, in many species, two CIRCE motifs were detected in the intergenic region between these divergent TUs; i.e. one close to the *groES/EL* TU, and the second close to the protease-encoding gene. Nevertheless, this finding does not necessarily indicate that HrcA regulates the transcription of this protease, since multiple *cis*-acting elements are often found to occur in front of a regulated TU.

**Ribosomal proteins.** Several different motifs were identified upstream of COTs encoding ribosomal proteins. These motifs were conserved among equivalent genes in different species, but the motifs identified in COTs with different ribosomal proteins were not clustered by COMPASS. It has been shown that ribosomal protein L4 expression is auto-regulated (30) while another study has predicted that this occurs for 43 other ribosomal proteins (31). These findings suggest that TUs encoding different ribosomal proteins are auto-regulated by the ribosomal protein they encode, each recognizing different upstream motifs. The *cis*-element predictions presented here are in good agreement with this mode of regulation of this class of proteins.

### Regulatory motif variations and group-specific motifs

Although many motifs were found to be present in both species sets, the regulatory network of these organisms can still be different. In several cases, the same motif was found to regulate different genes in the different organisms. Most of these differences in the prediction of the regulatory elements between the two species sets were caused by differences in the gene content or organization of the sets, rather than by differences in regulatory mechanisms. In general, it can be stated that if a gene is present in two different organisms, it is regulated by the same regulator. Some differences in motif prediction are displayed in Table 2 and discussed below.

**Table 2.** Selected regulatory motifs predicted by both species sets

LAB set	BAC set	Description	Regulated genes in <i>L.plantarum</i>	c
1		T-box, element in front of different tRNA synthetases and related genes	Many(>10) different tRNA synthetases. Amino acid biosynthesis and transport genes	n/a <sup>a</sup>
2		CIRCE element, cis-acting element of HrcA, involved in heat shock response	<i>groES-groEL, hrcA-grpE-dnaK, lp0726(membrane-bound protease)</i>	0.70
3		PyrR element, cis-acting element of PyrR, involved in pyrimidine biosynthesis.	<i>pyrR1, pyrP, pyrR2-pyrAA2-pyrAB2</i>	0.36
4		LexA element, cis-acting element of DinR, regulator of the SOS regulon.	<i>lexA, parC-parE</i>	0.61
5		RFN element, DNA element, regulates genes involved in biosynthesis and transport of riboflavin	<i>ribA-ribB-ribH, lp1887(transporter)</i>	n/a <sup>b</sup>
6		Conserved element in front of peptide release factor 2/B. Regulation of proteins involved in translation	<i>prfB</i>	n/a <sup>c</sup>
7		Several different boxes found upstream of genes encoding ribosomal proteins; presumably auto-regulatory sites	<i>rplM-rpsI</i>  <i>infC-rpml-rplT</i>  <i>rplK</i>  <i>rplJ-rplL</i>  <i>rplU-lp1593-rpmA</i>	n/a <sup>c</sup>  n/a <sup>c</sup>  n/a <sup>c</sup>  n/a <sup>c</sup>
8		CRE-box, binding site for CcpA, the general catabolite control protein	Several PTS systems and other genes involved in sugar metabolism.	0.40
9		dnaA-box, regulates genes involved in chromosome replication	<i>dnaA, dnaN, lp0045</i>	0.70
10	Not identified	S-box, regulates methionine biosynthesis genes in many Gram-positive organisms.	None	n/a <sup>b</sup>
11		THI element, involved in thiamine biosynthesis. Found in different COTs	<i>lp0217-lp0218-lp0219, thiM-thiD-thiE-lp0116- lp0117</i>	0.53
12	Not identified	FUR, regulates the uptake of Fe	None	n/a <sup>b</sup>
13	Not identified	Stress response	<i>lp2521, lp3215, lp3441-lp3442, lp3128, lp2807, lp0124, lp0433</i>	0.64
14		Catabolite availability response	<i>lp2813, lp3422, lp3553-araA-araD-araB-araT, lp3591-rhaD-rhaA-lp3594-rhaB-lp3596</i>	0.52

For a complete list, see the Supplementary Data. The regulated genes shown are those found in *L.plantarum*; those in other species can be found in Supplementary Data. A dash between genes signifies the same TU, while a comma separates different TUs. Some of the motifs (10, 12 and 13) were not found for COTs using the LAB set. Nevertheless, occurrences of these motifs could still be found in LAB set species using MAST. n/a, not applicable.

<sup>a</sup>Since T-boxes have different specificities, depending on their specifier codons, genes regulated by a T-box are not co-regulated.

<sup>b</sup>Genes not in the dataset.

<sup>c</sup>If only one TU is found to have the conserved cis-acting element, no correlations for the regulon can be calculated.

**Table 3.** Predicted LAB-specific motifs

LAB	Description	Regulated genes in <i>L.plantarum</i>
1	CopR binding site, found in front of <i>copAB</i> genes for copper transport in most <i>Lactobacillaceae</i>	<i>copR</i> , <i>copA</i> , <i>copB-bsh3</i>
2	Transcriptional processes	<i>rpoB-rpoC</i>
3	Teichoic acid biosynthesis	<i>dltD-dltC1-dltB-dltA-dltX-pbpX2</i>
4	Cold shock response	<i>cspC</i> , <i>cspL</i>
5	Unknown. Conserved TU in all LAB-set species	<i>lp0779 – lp0781</i>
6	Translation. Hits were observed in other LAB genomes, in front of a TU containing <i>prfB</i>	None
7	Unknown, all hypothetical genes	<i>lp2178</i>
8	Unknown, only one (hypothetical) gene	<i>lp0045</i>
9	Hits were observed in other LAB genomes in front of a TU containing glucosidases and glycosyltransferases	None
10	Transcriptional processes	<i>dnaN</i>
11	Unknown, unclear relation between t-RNA ligases and other genes	<i>ileS</i> , <i>argS</i> , <i>mesJ-hprT-fisH</i>
12	T-box like, regulates leucine tRNA synthetase genes	<i>leuS</i>
13	T-box like, regulates aspartate and histidine tRNA synthetase genes	<i>aspS</i> , <i>hisS</i>
14	T-box like, regulates glycine tRNA synthetase genes	<i>glyS</i>
15	T-box like, regulates aspartate-ammonia ligase and asparagine tRNA synthetase (possibly one contiguous large regulatory element)	<i>asnS1-asnA</i>

**Thiamine and Riboflavin biosynthesis.** *Lactobacillaceae* and many other Gram-positive bacteria contain genes for the biosynthesis of the vitamins thiamine and riboflavin. These biosynthetic pathways are regulated by riboswitches, which are structural elements in the untranslated upstream sequence in the corresponding mRNA that form a binding pocket for a metabolite that regulates expression of that gene. In case of thiamine, biosynthesis is regulated via the so-called THI

element that can bind thiamine itself (32). Under thiamine-limiting conditions no thiamine is bound to the THI element allowing three-dimensional RNA structure formation that promotes transcription of the mRNA. Excess thiamine leads to folding of a transcription terminator loop in the untranslated upstream region due to thiamine binding to the THI element, resulting in premature transcription termination. For riboflavin biosynthesis, a similar mechanism has been described, where flavin mononucleotide (FMN), that is a metabolic derivative of



riboflavin, inhibits the expression of the riboflavin biosynthesis genes (33,34).

MEME analysis identified the THI element in both species sets, but upstream of partially different COTs (Table 2, motif 11). In the LAB set, the motif was found twice, once in front of a TU encoding an ABC transporter (lp0217–lp0219) for an unknown substrate and once in front of a TU encoding thiamine biosynthesis genes. In the BAC set the motif was only found in a COT with TUs encoding thiamine biosynthesis. The MAST analysis showed that in three species of the BAC set the motif also occurs in the upstream region of a set of transporters with unknown substrate. The substrate of these transporters remains unclear, but it is tempting to speculate that this transporter is involved in thiamine or thiamine-precursor transport. Analogously, these transporters show high homology to the thiamine (and related substrates) transporters in *Salmonella typhimurium* (*thiBPQ*) and *Escherichia coli* (*sfuABC*) (35). However, the transporter amino acid sequence also displays similarity with some cation transporters of Gram-positive organisms, suggesting that its involvement in transport of cations that may act as cofactor during thiamine biosynthesis cannot be excluded. Interestingly, not all TUs preceded by the THI element were assigned to a COT, and were only identified through MAST analysis using the MEME-determined motif.

For riboflavin biosynthesis, riboswitch (RFN) elements were found upstream of all TUs encoding riboflavin biosynthesis genes (Table 2, motif 5). In addition, RFN elements precede a BAC-set COT encoding transporters for which recent experiments in *B.subtilis* show that it transports riboflavin (Jorg Vogel, personal communication). In the LAB-set, a COT with orthologous genes encoding the same transporter also contained the RFN element. Although riboflavin transporters and/or biosynthesis genes are not present in all species that were analyzed, no clear phylogenetic distribution was found for these genes. Species can either have both transporter and biosynthesis genes (*L.plantarum*, *P.pentaseceus*, *B.subtilis*, *S.aureus*), only one of the two (*L.johnsonii*, *L.brevis*, *L.delbrueckii*, *S.pneumoniae*, *E.faecalis*), or lack both systems (*L.monocytogenes*, *L.casei*). Interestingly, the RFN element was not found in front of all TUs encoding the presumed riboflavin transporter.

**Methionine biosynthesis.** The predicted *cis*-acting elements for COTs involved in methionine biosynthesis are also different in the two species sets. In the BAC set, a *cis*-acting element known as the S-box is detected (Table 2, motif 10). The S-box is a regulatory element to which S-adenosylmethionine can bind, leading to transcription attenuation (36,37). In the LAB set, a T-box is identified instead of this S-box upstream of methionine biosynthesis genes. Presumably, the unloaded methionine tRNA can bind to this T-box to induce translation of the methionine biosynthesis TU. These different mechanisms for regulating methionine biosynthesis, including their different phylogenetic distribution, have been described before (38). In analogy, MAST analysis, using the predicted T-box detects several methionine biosynthesis TUs in *L.plantarum* (Table 2, motif 1), while searches with the MEME derived S-box sequence from the BAC set does not detect significant hits in the upstream regions of TUs of *L.plantarum*.

### Non-LAB motifs

Only three predicted motifs of the BAC set were considered entirely absent in the LAB species set, since hits for these motifs were only found at most in a single species of the LAB set, but are highly represented (at least present in four organisms) in the BAC set. These three motifs precede TUs encoding, respectively, a DNA polymerase sliding-clamp subunit, ribosomal protein L11 and CTP synthase. In cases where a hit was found in one of the LAB set species, the hit was always in *L.plantarum*. The presence of the motif in *L.plantarum* could be a result of noise, due to the presence of *L.plantarum* in the initial BAC set, thereby generating a motif that will fit a (non-specific, but similar) upstream region of the *L.plantarum* TU. To remove this potential artifact a new MEME analysis was run without the upstream sequence of the *L.plantarum* TU. These new analyses identified motifs similar to those generated by the original MEME analysis. Nevertheless, in two out of three cases, the MAST searches with these motifs no longer gave a hit with the *L.plantarum* TU from the original COT. One motif still had a good MAST hit with the *L.plantarum* TU. The identified motif resembles a DnaA binding box, which was also found in front of other COTs. This motif appeared to be specific for BAC set species, including *L.plantarum*. Although this motif cannot be found when using the LAB set, it may still be important for the analysis of the regulatory network of *L.plantarum*. Surprisingly, the other identified DnaA boxes (Table 2, motif 9) were found to be shared among all used species (both BAC and LAB sets).

### LAB-specific motifs

Eighteen motifs (Table 3) were considered LAB-specific, i.e. not present in more than one of the BAC set species, but present in at least four LAB set species. If a LAB-specific motif was present in only one of the BAC set species, in all cases this was in *L.plantarum*. To clearly establish that these motifs were really LAB-specific, they were searched in additional *Lactobacillaceae* (*L.gasseri*, *Lactobacillus acidophilus*) and other LAB (*L.mesenteroides* and *O.oeni*) genomes. All 18 motifs were found to occur in at least two of these other species, and in most cases (95%) were located in the upstream region of a TU that resembles the TUs of the original COT. Searches with the 18 LAB-specific motifs in the intergenic regions of other publicly available genomes resulted only in false-positive hits (i.e. in upstream regions of non-related TUs), supporting the LAB-specificity of these *cis*-element-COT combinations.

Some of these 18 LAB-specific motifs were found in front of TUs with genes encoding clearly described functions but only in some cases a well-known regulatory motif. These include five tRNA-synthetases (for Gly, Ile, Leu, His, Asp), a tRNA methyltransferase, DNA polymerase III beta-subunit, RNA polymerase beta-subunit, peptide chain release factor, a lipopolysaccharide 1,2-glucosyltransferase, CopAB ATPases, cold shock proteins, and cell wall biosynthesis proteins (*dlt* TU). Four different LAB-specific motifs are found preceding a COT encoding aspartate-ammonia ligase. These motifs were found in the same order and with identical spacing regions between the motifs, suggesting that these motifs could act together as one large regulatory element, like a riboswitch.

**Copper transport.** All *Lactobacillaceae* share the same copper transporting ATPases (CopA, or, in *L.plantarum*, CopA and CopB). Cop genes are found in the genus of *Lactobacillaceae* and other LAB, such as *Lactococci* and *Streptococci*. In the LAB set, 1 COT was found to have the conserved binding site for CopR (Table 3, motif 1), the regulatory protein of the *copAB* genes (39). Some variation was found in the TU organization amongst the other species. Some species had one TU in which all *cop* genes are represented (*L.brevis*, *L.johnsonii*, *P.pentosaceus*), while other species encode the *cop* genes divided over two (*L.brevis*, *L.casei*) or even three (*L.plantarum*) different TUs. No *cop* genes were found in *Lactobacillus delbrueckii*, which is possibly due to the incompleteness of its currently available genome sequence. With exception of one of the *cop* genes in *L.plantarum*, the upstream regions of all TUs had a good hit with the predicted CopR binding site ( $P$ -value  $< 1 \times 10^{-11}$ ). Searches in species outside the initial LAB set showed that the motif is present in other LAB. Hits were found upstream of copper transporting ATPases in *Lactococcus lactis*, *Streptococcus agalactiae* and *Streptococcus thermophilus*. Notably, the motif was not conserved in other *Streptococcaceae* species, such as *Streptococcus pyogenes* and *S.pneumoniae*. This is especially remarkable since *S.thermophilus* and *S.agalactiae* are considered more distantly related to each other than to these other *Streptococci*. No hit with comparable  $P$ -value was found in the TUs of the BAC set. However, when searching for occurrences of the motif in other genomes, the CopR-binding site was found in front of genes annotated as penicillinase repressors in several *Bacillus cereus* and *Bacillus anthracis* strains. BLAST searches showed that these genes resemble the *copR* genes of the LAB genomes.

**Unknown functions.** Next to this well-described motif, several LAB-specific motifs were found that seem to be highly conserved among the different LAB species, including *Lactobacillaceae* (or related LAB), but for which the function remains unknown. To the best of our knowledge, none of these motifs have been described in literature to date. One example is motif 7 (Table 3), which is highly conserved in all LAB-set genomes (lowest  $P$ -value =  $5.3 \times 10^{-11}$ ), while no hits below threshold (of  $1.0 \times 10^{-5}$ ) can be found in the BAC set species. Nevertheless, the genes in the corresponding COT are found to have an ortholog in many of the BAC-set species. Moreover, searches in other available genomes (either publicly available, or accessible through the ERGO bioinformatics suite) revealed that these genes (encoding proteins of unknown function), as well as their relative order, appear widely conserved among prokaryotes. Nonetheless, the motif identified here appears to be uniquely present in LAB genomes. In addition to the initial LAB species set, hits were found in *L.gasseri* and *L.acidophilus*. Additional novel, LAB-specific motifs are listed in Table 3. Unfortunately, most of the predicted LAB-specific regulons consist of only one TU and can thus not be validated using the *L.plantarum* expression data from a single species like *L.plantarum* (see below).

### Regulon validation

To validate the predicted regulons in *L.plantarum*, expression correlations were calculated between genes that were part of the same regulon. For regulon validation, correlations of

genes within the same TU were discarded, since comparison of these genes would only validate TU prediction. Only genes with a high variance in expression ratio were used to reduce noise related to the small size of the test set (only 37 experiments). By applying these constraints, conclusions related to the accuracy of the phylogenetic footprinting could not be drawn for all predicted *L.plantarum* regulons. Nevertheless, for many of the predicted regulons a clear correlation of expression could still be observed, including several well-known *cis*-acting elements like CIRCE, LexA, DnaA and the THI element (Table 2, motif 2, 4, 9, 11). For the predicted regulons, the absolute correlation ( $|c|$ ) is shown, which is the mean of all absolute correlations for gene pairs in a regulon that do not share the same TU.

Microarray data analysis clearly established a highly correlated expression of *L.plantarum* TUs predicted to be encompassed within the *hrcA* regulon, including the TUs containing *groES/EL* and *hrcA*, *grpE* and *dnaK*. These TUs displayed a high expression correlation ( $c = 0.70$ ) in all experiments. Interestingly, the expression of gene lp0726, encoding a membrane-bound protease, located in the opposite transcriptional direction upstream of the *groES/EL* TU, also correlated with the *groES/EL* and *hrcA-grpE-dnaK* TUs of this regulon, albeit to a lesser extent ( $c = 0.60$  *groES/EL* TU,  $c = 0.35$  *hrcA* TU). This finding corroborates the functionality of the CIRCE element in the upstream region of this gene, and suggests a role of this protein in stress response. Overall, the mean correlation for the predicted CIRCE regulon is 0.59. Expression data analysis also supports the presence of THI elements in front of both the thiamine biosynthesis and transport encoding TUs in *L.plantarum*. Overall the correlation between these two different TUs is 0.53. Removal of one gene of the predicted regulon (lp0116) of *L.plantarum* that displays very poor correlation with the other genes of the regulon (including the ones in the same TU) increases the overall correlation to 0.61.

In addition to these well-known regulons, the expression data also validated several novel predicted regulons which have not been described in literature before. Two of these are shown in Table 2 (motif 13, 14). These identified regulons have an expression correlation of 0.64 and 0.52, respectively, which is in the same range as the well-known, literature-supported regulatory elements.

The first regulon consists of seven TUs (Table 2, motif 13), of which four showed to be suitable for validation. One of the TUs consists of two genes, while all others are monocistronic. Two of the encoded proteins have a predicted function in general stress response (lp3441 and lp3128), one is predicted to be a transcriptional regulator (lp2521) and the two others still have an unknown function (lp3215, lp3442). One of these genes (lp3442) encodes a protein with two conserved Interpro domains that are involved in binding and detoxification of heavy metals. This suggests that this regulon probably encodes a response to stress, possibly related to heavy metals. The *cis*-acting element preceding the TUs in the regulon was only identified by the BAC set, but hits were found in some of the LAB-set species, including *L.plantarum*.

Another regulon that could be validated with expression data was found by both species sets, but with a small difference in TU content. Both species sets predicted a regulon consisting of two large TUs (of five and six genes), both involved in

the breakdown of pentose sugars. Four out of five genes (*araA-D-T-B*) in the first TU have functions involved in the breakdown of arabinose, while three out of six genes (*rhaD-A-B*) in the second TU function in the breakdown of rhamnose. The other genes in these two TUs have either a general function in sugar breakdown (*maa3*) or transport (*lp3591*, *lp3596*), or are annotated as (conserved) genes with an unknown function (*lp3594*). In addition to these large TUs, both sets each predict an additional monocistronic TU containing a gene of unknown function. Addition of these TUs to the regulon does not influence the overall correlation of the complete regulon (from 0.51 to 0.52 in both cases).

## DISCUSSION

*L.plantarum* *cis*-acting elements were predicted based on phylogenetic foot-printing. In contrast to previous phylogenetic foot-printing studies, two different species sets were used that had different evolutionary distance to *L.plantarum*. In both species sets, numerous possible regulatory motifs were detected. Although there was significant overlap between the conserved regulatory motifs detected in each species set, several specific differences were found. Many approaches have been used to identify conserved *cis*-acting elements between species. In some studies, large sets of evolutionary quite distant species have been compared (11), while others compare only a few, closely related species (12). The present study shows that depending on the species being compared, different *cis*-acting elements will be predicted. Species sets with large evolutionary distances between the individual species will predict evolutionary highly conserved regulatory mechanisms, such as T-boxes and stress response regulation (CIRCE, LexA). The corresponding regulatory processes can thus be classified as generally conserved in many microorganisms. On the other hand, comparing species sets of closer related organisms will reveal more information about genus-specific regulatory mechanisms. As an example, the *cis*-acting element for the transcription factor CopR can only be found when performing a phylogenetic footprint analysis with a species set containing *Lactobacillus* species. When analyzing a set with Gram-positive organisms, the CopR binding site will not be found, as only a small proportion of the species analyzed will have a CopR-binding site.

Motif predictions performed with the LAB set identified 18 LAB-specific *cis*-acting elements, of which 14 can be considered unique (not part of a large *cis*-acting element together with other identified *cis*-acting elements) and not described in literature before. Sixteen LAB-specific motifs are present in *L.plantarum* (Table 3). Some of these elements seem to regulate specific biochemical pathways (*dltX*, involved in teichoic acid biosynthesis; Table 3, motif 3), while for other elements the function remains unknown (Table 3, motif 8).

It can be concluded that using different species in the phylogenetic footprinting analysis leads to differences in predicted regulatory motifs. We have shown that both species sets make a different contribution to the prediction of the regulatory network of *L.plantarum*. Each of the species sets predict *cis*-acting elements that cannot be found with the other species set. In many cases these differences are caused by differences in genome content; if genes are conserved between different

species, the regulatory element is conserved as well. Nevertheless, these motifs can still predict different TUs to be part of the regulon.

Microarray data from different experiments were used to validate the regulon predictions. By comparing the expression profiles of genes within a predicted regulon, conclusions could be drawn on the success rate of the prediction. However, since only a limited amount of microarray data are available for *L.plantarum*, validation is still limited to a few examples. With the growth of the number of microarray experiments more predictions will potentially be validated, leading to new insight in the regulatory network of *L.plantarum*.

In conclusion, when predicting the regulatory network in a genome of interest by phylogenetic footprinting, it is essential to select species that are evolutionary closely related, but in addition have comparable gene content. The data generated in this analysis can be of a great help to future microarray experiments. *Cis*-acting element predictions can find subsets of co-regulated genes within the larger set of co-expressed genes. This will help to elucidate the status of the regulatory network under the tested conditions and give hints to which environmental signals the organism responds.

## ACKNOWLEDGEMENTS

Thanks to Douwe Molenaar, Mariela Hebben and Arno Wegkamp for sharing and discussing their micro-array data. Thanks to Greer Wilson for intensive reading of the manuscript and critical discussions. Financial support was received from the IOP Genomics Programme grant no IGE1018. Funding to pay the Open Access publication charges for this article was provided by the Wageningen Centre for Food Sciences.

*Conflict of interest statement.* None declared.

## REFERENCES

- Cao, M., Kobel, P.A., Morshedi, M.M., Wu, M.F., Paddon, C. and Helmann, J.D. (2002) Defining the *Bacillus subtilis* sigma(W) regulon: a comparative analysis of promoter consensus search, run-off transcription/microarray analysis (ROMA), and transcriptional profiling approaches. *J. Mol. Biol.*, **316**, 443–457.
- Conway, T. and Schoolnik, G.K. (2003) Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. *Mol. Microbiol.*, **47**, 879–889.
- van Nimwegen, E., Zavolan, M., Rajewsky, N. and Siggia, E.D. (2002) Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics. *Proc. Natl Acad. Sci. USA*, **99**, 7323–7328.
- Mwangi, M.M. and Siggia, E.D. (2003) Genome wide identification of regulatory motifs in *Bacillus subtilis*. *BMC Bioinformatics*, **4**, 18.
- Thompson, W., Rouchka, E.C. and Lawrence, C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.
- Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell Syst. Mol. Biol.*, **2**, 28–36.
- Snel, B., van Noort, V. and Huynen, M.A. (2004) Gene co-regulation is highly conserved in the evolution of eukaryotes and prokaryotes. *Nucleic Acids Res.*, **32**, 4725–4731.
- Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.

9. McGuire, A.M., Hughes, J.D. and Church, G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.*, **10**, 744–757.
10. McCue, L.A., Thompson, W., Carmack, C.S. and Lawrence, C.E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.*, **12**, 1523–1532.
11. Alkema, W.B., Lenhard, B. and Wasserman, W.W. (2004) Regulog analysis: detection of conserved regulatory networks across bacteria: application to *Staphylococcus aureus*. *Genome Res.*, **14**, 1362–1373.
12. Yan, B., Methe, B.A., Lovley, D.R. and Krushkal, J. (2004) Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family Geobacteraceae. *J. Theor. Biol.*, **230**, 133–144.
13. Vaughan, E.E., de Vries, M.C., Zoetendal, E.G., Ben-Amor, K., Akkermans, A.D. and de Vos, W.M. (2002) The intestinal LABs. *Antonie Van Leeuwenhoek*, **82**, 341–352.
14. Davidson, B.E., Kordias, N., Dobos, M. and Hillier, A.J. (1996) Genomic organization of lactic acid bacteria. *Antonie Van Leeuwenhoek*, **70**, 161–183.
15. Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Turchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W. *et al.* (2003) Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc. Natl Acad. Sci. USA*, **100**, 1990–1995.
16. Boekhorst, J., Siezen, R.J., Zwahlen, M.C., Vilanova, D., Pridmore, R.D., Mercenier, A., Kleerebezem, M., de Vos, W.M., Brussow, H. and Desiere, F. (2004) The complete genomes of *Lactobacillus plantarum* and *Lactobacillus johnsonii* reveal extensive differences in chromosome organization and gene content. *Microbiology*, **150**, 3601–3611.
17. Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, E.Jr, Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I. *et al.* (2003) The ERGO genome analysis and discovery system. *Nucleic Acids Res.*, **31**, 164–171.
18. Moreno-Hagelsieb, G. and Collado-Vides, J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**, S329–S336.
19. Ermolaeva, M.D., Khalak, H.G., White, O., Smith, H.O. and Salzberg, S.L. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.*, **301**, 27–33.
20. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
21. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
22. Sadreyev, R. and Grishin, N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
23. Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.*, **32**, D75–D77.
24. Bailey, T.L. and Gribskov, M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
25. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
26. Grundy, F.J. and Henkin, T.M. (2003) The T box and S box transcription termination control systems. *Front Biosci.*, **8**, d20–31.
27. Zuber, U. and Schumann, W. (1994) CIRCE, a novel heat shock element involved in regulation of heat shock operon dnaK of *Bacillus subtilis*. *J. Bacteriol.*, **176**, 1359–1363.
28. Hecker, M., Schumann, W. and Volker, U. (1996) Heat-shock and general stress response in *Bacillus subtilis*. *Mol. Microbiol.*, **19**, 417–428.
29. Arnau, J., Sorensen, K.I., Appel, K.F., Vogensen, F.K. and Hammer, K. (1996) Analysis of heat shock gene expression in *Lactococcus lactis* MG1363. *Microbiology*, **142**, 1685–1691.
30. Stelzl, U., Zengel, J.M., Tovbina, M., Walker, M., Nierhaus, K.H., Lindahl, L. and Patel, D.J. (2003) RNA-structural mimicry in *Escherichia coli* ribosomal protein L4-dependent regulation of the S10 operon. *J. Biol. Chem.*, **278**, 28237–28245.
31. Abreu-Goodger, C., Ontiveros-Palacios, N., Ciria, R. and Merino, E. (2004) Conserved regulatory motifs in bacteria: riboswitches and beyond. *Trends Genet.*, **20**, 475–479.
32. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. and Gelfand, M.S. (2002) Comparative genomics of thiamin biosynthesis in prokaryotes. New genes and regulatory mechanisms. *J. Biol. Chem.*, **277**, 48949–48959.
33. Mack, M., van Loon, A.P. and Hohmann, H.P. (1998) Regulation of riboflavin biosynthesis in *Bacillus subtilis* is affected by the activity of the flavokinase/flavin adenine dinucleotide synthetase encoded by ribC. *J. Bacteriol.*, **180**, 950–955.
34. Lee, J.M., Zhang, S., Saha, S., Santa Anna, S., Jiang, C. and Perkins, J. (2001) RNA expression analysis using an antisense *Bacillus subtilis* genome array. *J. Bacteriol.*, **183**, 7371–7380.
35. Webb, E., Claas, K. and Downs, D. (1998) thiBPQ encodes an ABC transporter required for transport of thiamine and thiamine pyrophosphate in *Salmonella typhimurium*. *J. Biol. Chem.*, **273**, 8946–8950.
36. Grundy, F.J. and Henkin, T.M. (1998) The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol. Microbiol.*, **30**, 737–749.
37. McDaniel, B.A., Grundy, F.J., Artsimovitch, I. and Henkin, T.M. (2003) Transcription termination control of the S box system: direct measurement of S-adenosylmethionine by the leader RNA. *Proc. Natl Acad. Sci. USA*, **100**, 3083–3088.
38. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. and Gelfand, M.S. (2004) Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. *Nucleic Acids Res.*, **32**, 3340–3353.
39. Mills, S.D., Lim, C.K. and Cooksey, D.A. (1994) Purification and characterization of CopR, a transcriptional activator protein that binds to a conserved domain (cop box) in copper-inducible promoters of *Pseudomonas syringae*. *Mol. Gen. Genet.*, **244**, 341–351.