Unraveling the regulatory network of *Lactobacillus plantarum* WCFS1

Michiel W.W. Wels

Promotoren:
Prof. Dr. Willem M. de Vos
Hoogleraar Microbiologie
Wageningen Universiteit

Prof. Dr. Roland J. Siezen
Hoogleraar Bacterial Genomics
Radboud Universiteit Nijmegen

Co-Promotor:
Prof. Dr. Michiel Kleerebezem
Hoogleraar Bacterial Metagenomics
Wageningen Universiteit

Promotiecommissie:

Prof. Dr. T. Abee
Wageningen Universiteit

Prof. Dr. J.A.M. Leunissen
Wageningen Universiteit

Prof. Dr. O.P. Kuipers
Universiteit Groningen

Prof. Dr. M.S. Gelfand
Moscow University, Russia

Dit onderzoek is uitgevoerd binnen de onderzoeksschool VLAG.

Unraveling the regulatory network of *Lactobacillus plantarum* WCFS1

Michiel W.W. Wels

Proefschrift

ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit
Prof. dr. M.J. Kropff
in het openbaar te verdedigen
op maandag 7 januari 2008
des namiddags te vier uur in de Aula

Michiel WW Wels Unraveling the regulatory network of *Lactobacillus plantarum*
WCFS1

PhD thesis. Wageningen University and Research Centre, The Netherlands, 2007.
With summary in Dutch

ISBN: 978-90-8504-752-0

Table of Contents

Supplementary material: www.cmbi.ru.nl/~mwels/Thesis

Abstract

The ability of *Lactobacillus plantarum* to adapt to various environmental conditions, even the variable, complex and competitive conditions of the mammalian intestinal tract, makes it an interesting subject for studying the mechanisms underlying niche-specific adaptation. The focus of this thesis was to unravel the regulatory network in *L. plantarum* using different bioinformatics tools. In many cases the generated hypotheses were validated using evidence gathered from genomics experiments. The regulatory network of *L. plantarum* was analyzed using four different approaches; one based on the annotation of the regulatory proteins encoded on the genome, while the three other methods predict conserved *cis*-regulatory elements in the sequences upstream of encodes genes or Transcriptional Units (TUs). In the first method, putative locally-acting regulons were predicted using genome context conservation and complete genome hybridization (CGH) data. Comparison with genomes of other lactic acid bacteria revealed that *L. plantarum* has the highest relative fraction of its predicted proteome assigned to regulatory proteins and indicated that these differences can generally be related to regulatory protein families controlling the expression of genes related to adaptation to different or changing environments. The other three methods predicted conserved *cis*-acting regulatory elements, using different sources of information to initiate the search: i) the annotated genome sequence of *L. plantarum* WCFS1, ii) the annotated genome sequences of different sets of related bacterial species, and iii) multiple transcriptome datasets from a variety of experiments performed with *L. plantarum*. Two of the regulatory motifs found in these large-scale analyses were studied in more detail. A total of 24 highly conserved, genetically linked motifs (8 – 34 nt) that form multiple mosaic *L. plantarum* supermotifs (LPSM) were found in the genome sequence of *L. plantarum*. These LPSMs appear to be unique for *L. plantarum* but were found to be conserved among different *L. plantarum* strains. Although the function of these LPSMs remains unknown, transcriptome data suggested regulation of the expression in experiments comparing *L. plantarum* WCFS1 wild-type with a strain overexpressing endogenous genes from an expression plasmid. Phylogenetic footprinting, motif searching and RNA structure prediction procedures were employed to analyze the presence of T-box elements and their specifier codons in bacteria. The specifier codon of the various T-box elements was used to improve the functional annotation of approximately 125 genes in different bacterial genomes, including many genes that are notoriously difficult to annotate on basis of sequence similarity like amino acid transporters.

CHAPTER 1

General introduction

Acknowledgements

Thanks to Erwin Zoetendal and Mirjana Rajilic-Stojanovic for their help in creating an overview of the organization of the *Firmicutes* phylum.

The capacity to adapt to environmental conditions is crucial for organisms to survive and prosper in different ecological niches. Highly versatile and flexible bacteria are known to contain a large repertoire of genetic potential that enables adaptive variations of the expressed biochemical pathways and cellular functions, allowing the organism to cope with the different environmental situations it encounters. One of the major challenges in understanding the adaptation of an (microbial) organism is the unraveling of the gene-regulatory circuits that control transcription of subsets of genes, and to understand the underlying mechanisms of transcription regulation. An environmental change may lead to activation of genes required to cope with the altered condition, while repressing other genes that do not fulfill an important function anymore in the novel environment, leading to optimal tuning of the expression pattern of the genetic potential available in the cell. This introduction will provide an overview of the mechanisms to regulate the expression of the genes involved in these adaptational responses and general characteristics of *Lactobacillus plantarum*, the model-organism used in this thesis.

Firmicutes

Firmicutes are a phylum of bacteria (Figure 1) classified as Gram-positive on basis of their cell wall structure. Originally the *Firmicutes* were taken to include all Gram-positive bacteria, but on basis of 16S rRNA sequence conservation they are restricted to a core group of related organisms, called the low G+C group. This group includes the *Mollicutes* although they do not respond to Gram staining as they lack a cell wall (1). Other Gram-positive bacteria (the high G+C group) are nowadays classified as *Actinobacteria*. *Firmicutes* are predominantly found as round cells, called cocci (e.g. *Streptococcus sp.*), or rod-shaped forms (e.g. *Bacillus sp.*). Firmicutes also include many vaguely described bacterial groups such as *Clostridium* and *Eubacterium sp*. Several *Firmicutes* – e.g. *Clostridium* and *Bacillus* – are known to produce endospores, which are resistant to desiccation and can survive extreme conditions. *Firmicutes* are found in various environments, and some are notable pathogens. In addition, they play an important role in the production as well as spoilage of various food products (2). There is currently no methodology, other than 16S rRNA sequencing, to assign a bacterium as belonging to the *Firmicutes*. This is because the phylum is highly diverse in phenotypic characteristics due to promiscusious plasmid exchange across species and genera of this phylum. The *Firmicutes* phylum is divided into three orders: the *Clostridia*, which are anaerobes, the *Bacilli*, which are obligate or facultative aerobes, and *Mollicutes*. In phylogenetic trees the first two groups show up as paraphyletic or polyphyletic, as do their main genera, *Clostridium* and *Bacillus*.

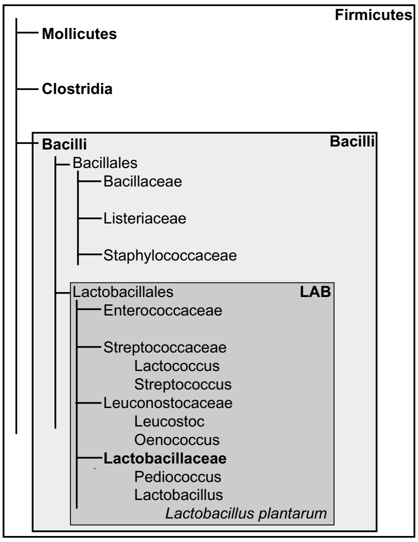


Figure 1: Partial schematic overview of the division of *Firmicutes* displaying the position of *L. plantarum* within this phylum.

Lactic acid bacteria (LAB)

Lactic acid bacteria (LAB) make up a large but relatively homogenous subgroup within the *Bacilli* (Figure 1). LAB are encountered in a large number of different environments and are well-known for their application in different food fermentation processes (2). In addition, several LAB have been shown to contribute to intestinal health and are therefore added to different probiotic food products (3).

With the development of high-throughput sequencing techniques, the genomes of a large number of LAB have been sequenced (4). The sequenced species range from the probiotic species *Lactobacillus acidophilus* and *Lactobacillus salivarius*, the industrially relevant yoghurt bacteria *Streptococcus thermophilus* and *Lactobacillus delbrueckii* to the pathogenic *Enterococcus faecalis* and *Streptococcus pneumoniae*. The availability of a large number of sequenced species and strains opens the possibility to do large-scale comparative genomics studies on the evolution of LAB (5,6).

Lactobacillus plantarum

Lactobacillus plantarum WCFS1 was the first *Lactobacillus* species for which the genome was completely sequenced (7). The genome is among the largest known among LAB (3.3 Mb) and is therefore a good candidate to act as a model organism for other *Lactobacilli*. *L. plantarum* is found in a variety of different niches, including the human gastrointestinal tract (GI-tract) (8). Its function in the intestine is still unclear, but research has pointed out that the presence of some *L. plantarum* strains in the human GI-tract can promote health. Several strains are marketed as a probiotic (e.g. *L. plantarum* 299v (9)) and have proven to be useful in the prevention of cardiovascular disease in smokers (10), the prevention of infection by pathogens through production of inhibitory factors or competition for specific niches and nutrients (11), decreasing symptoms in Irritable Bowel Syndrome (12), reducing both LDL-cholesterol and fibrinogen levels in blood (13) and many more (for a recent review see (14)). Next to its presence in the GI-tract, *L. plantarum* is encountered in variety of fermentation processes involving plant-derived raw materials, including olives (15,16), sauerkraut (17), cassave (18), but also as a predominant organism in silage processes (19). In addition to these plant material fermentations, *L. plantarum* can also be found in a variety of fermentation processes of other raw materials like milk (20-22), or meat (23-25), albeit that their occurrence is less frequent in these latter niches. Its ability to survive in a large number of different environments and metabolize a large variety of different substrates makes *L. plantarum* a suitable model organism to study the metabolic capacities and interactions with the environment. The recent development of different molecular toolboxes and genomics platforms has increased the level of detail in which *L. plantarum* can be studied (26,27).

Transcription regulation

The ability of *L. plantarum* to adapt to various environmental conditions, even the variable, complex and competitive conditions of the mammalian intestinal tract, makes it an interesting subject for studying the mechanisms underlying niche-specific adaptation. One of the major topics in understanding the adaptation of a (microbial) organism is the unraveling of the gene-regulatory circuits that control transcription of subsets of genes, and to understand the underlying mechanism of transcription regulation. An environmental change may lead to activation of genes required to cope with the altered condition, while repressing other genes that do not fulfill an important function anymore in the novel environment, overall leading to optimal tuning of the expression pattern of the genetic potential available in the cell.

Coregulation - TUs and regulons

In order to react to a changing environment, an organism will often require the presence of different genes that cooperate in the same process. To successfully adapt, the 'collaborating' gene-products need to be present at the same time and in the right concentrations. There are two ways to organize co-expression of two genes: i) co-transcription, transcribing the genes within a single transcriptional unit (TU), or ii) co-regulation, controlling the expression of both genes by concerted control of their expression, predominantly achieved by shared regulatory signals for both genes (regulator binding sites; i.e. *cis*-acting elements) that involve control by the same transcription factor (TF, *trans*-acting factor).

Genes under regulation of the same regulator are called a regulon. Regulation of expression of a set of genes often occurs in a local manner; the TF and its regulon are located in close vicinity of each other. As an alternative, and/or in addition to local regulation, genes can be regulated through more global mechanisms, involving global TFs that control the expression of many genes that may be located at very distant positions in the genome (28). These global TFs often code for a protein that is present in the cell in high concentrations in order to be able to bind to the DNA at many different sites simultaneously. Most organisms are known to contain less than 10 of these global TFs (TFs regulating the expression of more than 20 genes) (29).

Mechanisms of regulation

In bacteria, there are two major ways in which the level of transcription of genes is controlled. The first is constitutive expression. Constitutive expression depends on the structure of the promoter, a sequence region that is located at -90 to +20 around the transcription start point (TSP) (Figure 2A). The core sequence of the promoter is a nucleotide sequence that is recognized and bound by the sigma subunit of the RNA polymerase. Next to this sigma factor binding site, several other regions are known to affect the level of transcription. The region upstream of the sigma factor binding site (-40 to -70 upstream of the TSP) is known to be AT-rich due to conserved A and T tracts called UP elements (30-32). These UP elements assist in binding of the RNA polymerase by providing a docking site for the carboxyl-terminal domains of the α subunits of the RNA polymerase (32-34). In addition to the UP elements, the early transcribed region (+1 - +20 downstream of the TSP) can also influence transcription efficiency by influencing promoter escape (35).

The basal level of transcription depends for the largest part on variations in the sigma factor binding site sequence. The most commonly used sigma factor, SigA (or sigma 42, sigma70, RpoD), recognizes a sequence located around –35 and –10 nt upstream of the TSP (consensus TTGACA and TATAAT (30)). SigA is also known as the housekeeping sigma factor and is involved in control of the basal expression of a large fraction of all genes; e.g. 642 genes have experimentally been shown to be transcribed by the SigA containing RNA-polymerase holo-enzyme (sigma70) in *Escherichia coli* (RegulonDB, 16th of August 2007 (36)). Next to these common consensus sequences of the SigA binding site, an "extended" -10

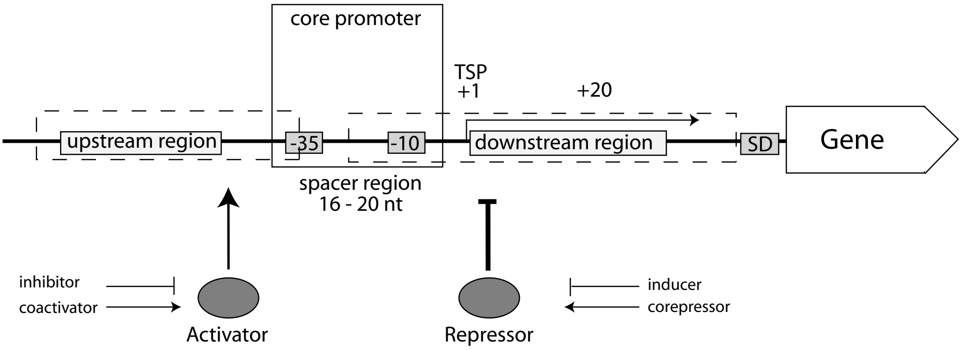


Figure 2: Structure of the bacterial promoter.
The core promoter is defined by the sequences that are directly bound by the sigma factor (in this case the -35 and -10 region or the "extended" -10 region of SigA) and the spacer region in between. The upstream region is known to contain AT-rich sequences that favor binding of the alpha subunit of the RNA polymerase. The downstream region is not generally conserved but can contain sequences that influence promoter strength and promoter escape efficiency. Just in front of the gene, the Shine-Dalgarno (SD) sequence is located. This element is not considered to be part of the promoter but assists in binding of the ribosome to the transcribed mRNA. Transcription factors (TFs) can bind at different locations in the promoter (indicated by the dashed lines) and can have positive (activator) or negative (repressor) effect on transcription. Intracellular signals (e.g. metabolites, peptides, small molecules) can alter the activity of the TF both positive (corepressor, coactivator) as well as negative (inhibitor, inducer).

region was described in *Bacillus subtilis* (30) (consensus TGnTATAAT). Although this "extended" -10 region was later also described in Gram-negative organisms (e.g. *E. coli* (37)), it is not as abundantly present as in Gram-positive organisms. Promoters that contain this extended -10 region were shown to be less dependant on a well-conserved -35 region (37). In addition, studies in *Lactococcus lactis* and *E. coli* have shown that the sequence and length of the spacer region in between the -35 and -10 region can also affect the strength of the promoter (38), but the properties of this spacer region could not be translated to distinct sequence or structural feature of this region.

In addition to the general SigA sigma factor, alternative sigma factors are known to exist. The alternative sigma factors are generally more dedicated for the transcription control of genes with a correlated function related to a single process. One of the better described examples in Gram-positive bacteria relates to the use of a complex regulatory circuit for the control of sporulation in *B. subtilis*, which includes the pivotal-regulatory function exerted by as many as five different sigma factors (for a review see (39)). The *L. plantarum* genome encodes three sigma factors; the general sigma factor (SigA) and two dedicated sigma factors (Sig54 and Sig30) (7). For the latter two it is assumed that they control the basal expression of only a limited number of genes.

The second mechanism to regulate the level of transcription of individual genes is regulatory control. These are means of regulation of gene expression, by which expression can respond to changes in the environment or biochemical needs of the organism. Regulatory control depends largely on transcription factors (TFs), which regulate transcription by binding on the DNA in most cases by interacting with the RNA polymerase (40-42). The DNA-binding sites of the TF are called regulatory binding sites or *cis*-acting elements. These elements are specific nucleotide sequences (or motifs) to which a regulatory element can bind in order to control the expression of a given gene.

There are two common strategies of transcription regulation found in bacteria; positive and negative control (Figure 2B). In positive control, binding of the TF (activator) on the *cis*-acting element will result in transcription of the downstream located TU, often by increasing the binding affinity of the RNA polymerase to the promoter or enhancing a subsequent rate-limiting step. Signals that aid in binding of the TF are called inducers or coactivators, while signals that relieve binding of the TF are known as inhibitors. The best known example of coactivated positive control is the expression of the *ara* operon by arabinose-activated AraC in *E. coli* (43). The best example of inhibition of positive control is the *leu* operon of *E. coli* that is normally activated by Lrp (Leucine responsive protein) but inhibited upon binding of leucine to Lrp (44). Although originally characterized in *E. coli*, regulatory proteins of the LRP family are found to regulate amino acid biosyntheses-related genes among many different prokaryotes (45).

In negative control, the TF (repressor) causes inhibition of transcription upon binding on the *cis*-acting element by blocking the binding of the RNA polymerase to the promoter or increasing promoter clearance (46-49). In this second case, signals that aid in binding of the TF to the *cis*-acting element are called corepressors while agents blocking TF-DNA interaction are called inducers. The best studied induced derepression system in prokaryotes is the *lac* operon of *E. coli* (50), which is normally repressed by LacI but transcribed when the inducer (lactose or IPTG) binds to LacI. The best known example of corepression is repression of transcription of the *trp* operon of *E.coli* (51) by binding of a complex of TrpR and tryptophan.

DNA binding of regulatory proteins

A common motif in DNA binding proteins, both in eukaryotes and prokaryotes, is the helix-turn-helix (HTH) motif. This motif was the first DNA-binding structure to be identified (52). The motif is made up by two α-helices, separated by a β-turn. This β-turn, together with the first α-helix, positions the second α-helix on the surface of the protein. This α-helix is now oriented in such a way that enables it to fit in the major groove of a DNA molecule. Therefore, the second α-helix is a recognition helix that makes contact with DNA to enable reading of the sequence by RNA polymerase. The HTH motif is only a minor part of the protein (about 20 amino acids). Some other parts of the protein connect to the DNA surface in order to establish correct positioning of the recognition helix. Also present in some regulatory proteins are binding sites that are specific for inducers. When the inducer binds at this site, the protein conformation changes to influence the activity of the operator-binding site. Nearly every prokaryotic protein that uses a HTH motif to bind DNA functions as a homodimer, with each motif binding one half-site of symmetry-related sequence motif.

The binding of the regulatory proteins to DNA occurs at highly specific positions. The regulatory proteins recognize a specific sequence (or motif) with their binding site. They can bind to this site by forming non-covalent contacts with chemical groups that are exposed within the major and minor groove of the double helix. Although the nucleotide bases are on the inside of the molecule, some of their chemical groups are accessible from outside the helix. This makes direct recognition of the nucleotide sequence by protein binding sites possible.

Non-protein mediated regulation

Next to protein-mediated regulation, other, more direct regulatory mechanisms occur in bacteria. Riboswitches are elements located in the upstream region of a TU that, as an untranslated RNA molecule, act to a specific signal without a protein intermediate. Interaction can cause either continuation or termination of transcription. Although many riboswitches are known to act upon binding of a metabolite (53), other interactions, like binding to a tRNA (to so-called T-boxes (54)) or acting on a change of temperature have also been described (for a review see (55)).

In addition to riboswitches, recent developments have also shown the presence of *trans*-acting regulatory RNA elements. The existence of regulatory RNA elements, that act *in trans* with mRNA molecules to regulate the expression of a certain gene was first shown to occur in Eukaryotes, but was soon after discovered in prokaryotes as well. Genome-wide studies in *E. coli* (56) and different *Archeae* (57,58) have shown that these organisms contain at least 100 different small RNAs, for which it is thought that many are involved in regulatory processes. Although the mechanisms underlying the exact mode of action are still unknown for most of these RNA elements, several publications have shown the power and elegance of these RNA-mediated regulatory mechanisms in controlling the expression of a variety of different genes (59-61).

Detection of conserved DNA elements

Regulatory proteins bind at specific DNA structures in the upstream region of the regulated gene. As most regulatory proteins function as homodimers, the recognized DNA sequences often reflect (incomplete) direct repeats or palindromes. In the past many efforts have been made to detect such conserved regulatory binding sites (62,63). In general there are two often-used strategies to come to a high quality prediction of regulatory binding sites, i.e. starting with a set of potentially coregulated genes (64-66), or on basis of phylogenetically related promoter regions (67-70) (Figure 3). In the first strategy it is assumed that coexpressed genes (e.g. identified in microarray experiments) are possibly regulated by the same regulatory protein and therefore share the same binding site in their promoter region. In the second strategy, called phylogenetic footprinting, it is assumed that the promoter regions of orthologous genes in different species are in itself orthologous and therefore share the same regulatory binding site. Both methods have been shown to be effective in the prediction of new regulatory binding sites in the promoter regions of genes (64-70).

The strategies described require algorithms to search for regulatory binding sites in a set of regulated DNA sequences. These algorithms can be divided into two different categories; deterministic and probabilistic (Table 1). In deterministic models searches often involve determining the presence of sequence strings

Table 1: Motif prediction algorithms

Algorithm	Advantages	Disadvantages	Exemplary tool
Deterministic	Fast	Consensus-based (needs a consensus sequence)	WEEDER (71)
Probabilistic (Expectation Maximization)	Flexible	Risk of getting stuck in a local optimum	MEME (73)
Probabilistic (Stochastic)	Global optimization	Stringent in settings	Gibb's Motif Sampler (85)

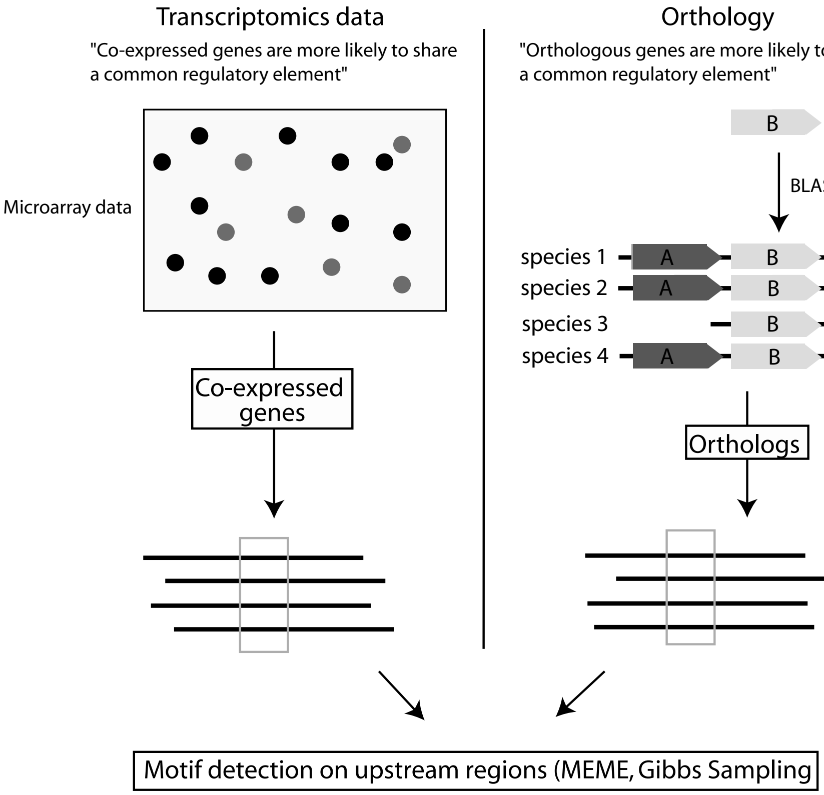

Transcriptomics data
"Co-expressed genes are more likely to share a common regulatory element"
Microarray data
Co-expressed genes
Orthology
"Orthologous genes are more likely to share a common regulatory element"
B
BLAST
species 1
species 2
species 3
species 4
A
B
C
D
Orthologs
Motif detection on upstream regions (MEME, Gibbs Sampling

Figure 3: Strategies for predicting cis-regulatory elements in bacteria.
The most commonly used methods use transcriptomics data or orthology. The upstream regions of coexpressed or orthologous genes are fed into motif-detection tools

or regular expressions (called "words") in a given set of DNA sequences. These methods are especially useful when based on a known consensus sequence. Comparison of the predicted binding site with the consensus sequence will allow scoring of statistical relevance (71,72).

Probabilistic models do not require a consensus sequence for detection of a possible binding site in a set of sequences. These methods are based upon local alignment of small regions within the submitted set of sequences. Probabilistic methods also come in two variants, Expectation Maximization (EM) and Gibb's Sampling. EM-based methods (73,74) are based upon local optimization of the alignment by increasing the statistical likelihood of a predicted binding site to not be randomly found. The biggest drawback of EM is the risk of getting stuck in a local optimum. This problem can be overcome by detecting multiple motifs in one run and selecting the statistically most significant motif. Gibb's Sampling (75,76) is a stochastic variant of EM and is therefore also suited to solve the risk of getting stuck in a local optimum. However this stochastic nature of the method requires the settings of these tools to be more stringent, for example by setting motif detection as a fixed width.

Regulatory networks

With the development of large-scale genomics techniques like transcriptomics and ChIP-chip experiments, different efforts to analyze the complete regulatory network of an organism were performed in well-known model organisms such as *E. coli* and yeast (77-82). Although these first analyses were based purely on experimental data, later efforts included knowledge gathered from literature and databases with curated information on regulatory networks to refine the predicted network (81,83,84). Application and translation of these networks to accelerate the network prediction in other, less well-studied organisms is a challenge still to be met in the future.

Outline of this thesis

The research presented in this thesis involves the characterization of the regulatory network of *L. plantarum*. In all cases, the genome sequence of the organism played a central role in the analysis, but was enriched using comparative genomics and/or transcriptomics data. With the application of several different techniques, various approaches were combined to get insight in the regulatory mechanisms underlying the interaction of *L. plantarum* with its environment.

Chapter 2 describes an in depth analysis of the full complement of regulatory proteins in *L. plantarum*. The genome encodes 206 regulatory proteins that were divided into different families based on COG and PFAM annotations. A comparison was made with other sequenced LAB genomes. *L. plantarum* was shown to have the largest number of regulatory proteins, both absolute as well as related to the genome size. This large number of regulatory proteins reflects the genome size and flexibility of *L. plantarum*.

Chapter 3 deals with the prediction of *cis*-acting elements on the genome, using all intergenic DNA sequences as an input for the motif prediction software. We identified the ribosome-binding site (Shine-Dalgarno sequence), various promoter elements and regulatory binding sites for a few globally acting regulators. The relation between gene start and promoter location was used as a validation criterion.

Chapter 4 discusses the identification of a large, repeated, intergenic DNA sequence encountered in the search for *cis*-acting elements. This *L. plantarum* supermotif (LPSM) was shown to occur 24 times on the chromosome and predicted to fold into a specific three-dimensional structure that resembles a cruciform. On basis of microarray data, it was shown that the LPSM is highly expressed upon the introduction of plasmids that overexpress endogenous gene products and most probably acts as a regulatory RNA molecule.

In Chapter 5 an alternative method for *cis*-acting element prediction was used. Using comparative genomics based on transcriptional unit conservation between different *Lactobacillaceae* and *Bacilli* we predicted a large number of regulatory elements in *L. plantarum*. Predictions were validated using correlated expression data gathered from a large amount of different microarray studies.

Chapter 6 describes the in-depth analysis of T-boxes, one of the motifs identified in chapter 5. T-box regulatory RNA elements, recognized by tRNAs, were searched for in all bacterial genomes and found to be widespread among all *Firmicutes*. Using sequence alignment and structure prediction, the tRNA specificity of the T-box was determined and used for functional annotation of the downstream-located genes.

In Chapter 7 a third method for identification of *cis*-acting elements is described. Correlated expression profiles gathered from microarray experiments are used as a source to identify potentially co-regulated genes. These analyses were especially successful in determining regulons that evolve at a high evolutionary rate. In addition, we could show that expression of a single gene is often dependant on the cross-talk between different regulatory proteins.

Chapter 8 summarizes the results and addresses some future prospects in the analysis of regulatory networks in bacteria. In addition we discuss the usefulness of the addition of genomics data gathered from *in vivo* gene expression studies, e.g. the gene expression of *L. plantarum* in the gastrointestinal tract of mice or men.

The Appendix describes the construction of **Correlex**, a web-accessible database containing microarray data of *L. plantarum* gathered from a large number of microarray experiments. This data can be used to study correlated gene expression and elucidate underlying regulatory properties.

References

1. **Dandekar, T., Snel, B., Schmidt, S., Lathe, W., Suyama, M., Huynen, M. and Bork, P. (2002)** Comparative genome analysis of the Mollicutes. *In Molecular Biology and Pathogenicity of Mycoplasmas. Kluwer, New York.*
2. **Holzapfel, W.H., Haberer, P., Geisen, R., Bjorkroth, J. and Schillinger, U. (2001)** Taxonomy and important features of probiotic microorganisms in food and nutrition. *Am J Clin Nutr, 73, 365S-373S.*
3. **Saxelin, M., Tynkkynen, S., Mattila-Sandholm, T. and de Vos, W.M. (2005)** Probiotic and other functional microbes: from markets to mechanisms. *Curr Opin Biotechnol, 16, 204-211.*
4. **Liu, M., van Enckevort, F.H. and Siezen, R.J. (2005)** Genome update: lactic acid bacteria genome sequencing is booming. *Microbiology, 151, 3811-3814.*
5. **Canchaya, C., Claesson, M.J., Fitzgerald, G.F., van Sinderen, D. and O'Toole, P.W. (2006)** Diversity of the genus *Lactobacillus* revealed by comparative genomics of five species. *Microbiology, 152, 3185-3196.*
6. **Makarova, K.S. and Koonin, E.V. (2007)** Evolutionary genomics of lactic acid bacteria. *J Bacteriol, 189, 1199-1208.*
7. **Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W. et al. (2003)** Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A, 100, 1990-1995.*
8. **Ahrne, S., Nobaek, S., Jeppsson, B., Adlerberth, I., Wold, A.E. and Molin, G. (1998)** The normal *Lactobacillus* flora of healthy human rectal and oral mucosa. *J Appl Microbiol, 85, 88-94.*
9. **Cunningham-Rundles, S., Ahrne, S., Bengmark, S., Johann-Liang, R., Marshall, F., Metakis, L., Califano, C., Dunn, A.M., Grassey, C., Hinds, G. et al. (2000)** Probiotics and immune response. *Am J Gastroenterol, 95, S22-25.*
10. **Naruszewicz, M., Johansson, M.L., Zapolska-Downar, D. and Bukowska, H. (2002)** Effect of *Lactobacillus plantarum* 299v on cardiovascular disease risk factors in smokers. *Am J Clin Nutr, 76, 1249-1255.*
11. **Kingamkono, R., Sjogren, E. and Svanberg, U. (1999)** Enteropathogenic bacteria in faecal swabs of young children fed on lactic acid-fermented cereal gruels. *Epidemiol Infect, 122, 23-32.*
12. **Niedzielin, K., Kordecki, H. and Birkenfeld, B. (2001)** A controlled, double-blind, randomized study on the efficacy of *Lactobacillus plantarum* 299V in patients with irritable bowel syndrome. *Eur J Gastroenterol Hepatol, 13, 1143-1147.*
13. **Bukowska, H., Pieczul-Mroz, J., Jastrzebska, M., Chelstowski, K. and Naruszewicz, M. (1998)** Decrease in fibrinogen and LDL-cholesterol levels upon supplementation of diet with *Lactobacillus plantarum* in subjects with moderately elevated cholesterol. *Atherosclerosis, 137, 437-438.*
14. **de Vries, M.C., Vaughan, E.E., Kleerebezem, M. and De Vos, W.M. (2006)** *Lactobacillus plantarum*—survival, functional and potential probiotic properties in the human intestinal tract. *International Dairy Journal, 16, 1018 - 1028.*
15. **Duran Quintana, M.C., Garcia Garcia, P. and Garrido Fernandez, A. (1999)** Establishment of conditions for green table olive fermentation at low temperature. *Int J Food Microbiol, 51, 133-143.*
16. **Randazzo, C.L., Restuccia, C., Romano, A.D. and Caggia, C. (2004)** *Lactobacillus casei*, dominant species in naturally fermented Sicilian green olives. *Int J Food Microbiol, 90, 9-14.*
17. **Stamer, J.R. (1983)** Lactic acid fermentation of cabbage and cucumbers. *H.J. Rehm and G. Reed (ed.), Biotechnology, vol. 5 Verlag Chemie, Weinheim, Germany.*
18. **Lei, V., Amoa-Awua, W.K. and Brimer, L. (1999)** Degradation of cyanogenic glycosides by *Lactobacillus plantarum* strains from spontaneous cassava fermentation and other microorganisms. *Int J Food Microbiol, 53, 169-184.*
19. **Klocke, M., Mundt, K., Idler, C., McEniry, J., O'Kiely, P. and Barth, S. (2006)** Monitoring *Lactobacillus plantarum* in grass silages with the aid of 16S rDNA-based quantitative real-time PCR assays. *Syst Appl Microbiol, 29, 49-58.*
20. **Ercolini, D., Hill, P.J. and Dodd, C.E. (2003)** Bacterial community structure and location in Stilton cheese. *Appl Environ Microbiol, 69, 3540-3548.*
21. **Manolopoulou, E., Sarantinopoulos, P., Zoidou, E., Aktypis, A., Moschopoulou, E., Kandarakis, I.G. and Anifantakis, E.M. (2003)** Evolution of microbial populations during traditional Feta cheese manufacture and ripening. *Int J Food Microbiol, 82, 153-161.*
22. **Baruzzi, F., Morea, M., Matarante, A. and Cocconcelli, P.S. (2000)** Changes in the *Lactobacillus*

community during Ricotta forte cheese natural fermentation. *J Appl Microbiol, 89, 807-814.*

23. **Cocolin, L., Manzano, M., Cantoni, C. and Comi, G. (2000)** Development of a rapid method for the identification of *Lactobacillus spp.* isolated from naturally fermented italian sausages using a polymerase chain reaction-temperature gradient gel electrophoresis. *Lett Appl Microbiol, 30, 126-129.*

24. **Enan, G., el-Essawy, A.A., Uyttendaele, M. and Debevere, J. (1996)** Antibacterial activity of *Lactobacillus plantarum* UG1 isolated from dry sausage: characterization, production and bactericidal action of plantaricin UG1. *Int J Food Microbiol, 30, 189-215.*

25. **Gevers, D., Danielsen, M., Huys, G. and Swings, J. (2003)** Molecular characterization of *tet(M)* genes in *Lactobacillus* isolates from different types of fermented dry sausage. *Appl Environ Microbiol, 69, 1270-1275.*

26. **Lambert, J.M., Bongers, R.S. and Kleerebezem, M. (2007)** Cre-lox-based system for multiple gene deletions and selectable-marker removal in *Lactobacillus plantarum*. *Appl Environ Microbiol, 73, 1126-1135.*

27. **Molenaar, D., Bringel, F., Schuren, F.H., de Vos, W.M., Siezen, R.J. and Kleerebezem, M. (2005)** Exploring *Lactobacillus plantarum* genome diversity by using microarrays. *J Bacteriol, 187, 6119-6127.*

28. **Martinez-Antonio, A. and Collado-Vides, J. (2003)** Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol, 6, 482-489.*

29. **Lozada-Chavez, I., Janga, S.C. and Collado-Vides, J. (2006)** Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res, 34, 3434-3445.*

30. **Helmann, J.D. (1995)** Compilation and analysis of *Bacillus subtilis* sigma A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res, 23, 2351-2360.*

31. **Ross, W., Aiyar, S.E., Salomon, J. and Gourse, R.L. (1998)** *Escherichia coli* promoters with UP elements of different strengths: modular structure of bacterial promoters. *J Bacteriol, 180, 5375-5383.*

32. **Ross, W., Gosink, K.K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K. and Gourse, R.L. (1993)** A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science, 262, 1407-1413.*

33. **Aiyar, S.E., Gourse, R.L. and Ross, W. (1998)** Upstream A-tracts increase bacterial promoter activity through interactions with the RNA polymerase alpha subunit. *Proc Natl Acad Sci U S A, 95, 14652-14657.*

34. **Estrem, S.T., Ross, W., Gaal, T., Chen, Z.W., Niu, W., Ebright, R.H. and Gourse, R.L. (1999)** Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes Dev, 13, 2134-2147.*

35. **Hsu, L.M. (2002)** Promoter clearance and escape in prokaryotes. *Biochim Biophys Acta, 1577, 191-207.*

36. **Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J. et al. (2006)** RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res, 34, D394-397.*

37. **Mitchell, J.E., Zheng, D., Busby, S.J. and Minchin, S.D. (2003)** Identification and analysis of ‘extended -10’ promoters in *Escherichia coli. Nucleic Acids Res, 31, 4689-4695.*

38. **Jensen, P.R. and Hammer, K. (1998)** The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters. *Appl Environ Microbiol, 64, 82-87.*

39. **Hilbert, D.W. and Piggot, P.J. (2004)** Compartmentalization of gene expression during *Bacillus subtilis* spore formation. *Microbiol Mol Biol Rev, 68, 234-262.*

40. **Mencia, M., Monsalve, M., Rojo, F. and Salas, M. (1996)** Transcription activation by phage phi29 protein p4 is mediated by interaction with the alpha subunit of *Bacillus subtilis* RNA polymerase. *Proc Natl Acad Sci U S A, 93, 6616-6620.*

41. **Mencia, M., Monsalve, M., Rojo, F. and Salas, M. (1998)** Substitution of the C-terminal domain of the *Escherichia coli* RNA polymerase alpha subunit by that from *Bacillus subtilis* makes the enzyme responsive to a Bacillus subtilis transcriptional activator. *J Mol Biol, 275, 177-185.*

42. **Monsalve, M., Mencia, M., Salas, M. and Rojo, F. (1996)** Protein p4 represses phage phi 29 A2c promoter by interacting with the alpha subunit of *Bacillus subtilis* RNA polymerase. *Proc Natl Acad Sci U S A, 93, 8913-8918.*

43. **Schleif, R. (2003)** AraC protein: a love-hate relationship. *Bioessays, 25, 274-282.*

44. **Newman, E.B., D'Ari, R. and Lin, R.T. (1992)** The leucine-Lrp regulon in *E. coli*: a global response in search of a raison d'etre. *Cell, 68, 617-619.*

45. **Brinkman, A.B., Ettema, T.J., de Vos, W.M. and van der Oost, J. (2003)** The Lrp family of transcriptional regulators. *Mol Microbiol, 48, 287-294.*

46. **Bird, T.H., Grimsley, J.K., Hoch, J.A. and Spiegelman, G.B. (1996)** The *Bacillus subtilis* response regulator Spo0A stimulates transcription of the *spoIIG* operon through modification of RNA polymerase promoter complexes. *J Mol Biol, 256, 436-448.*

47. **Cervin, M.A., Lewis, R.J., Brannigan, J.A. and Spiegelman, G.B. (1998)** The **Bacillus subtilis** regulator SinR inhibits *spoIIG* promoter transcription in vitro without displacing RNA polymerase. *Nucleic Acids Res, 26, 3806-3812.*

48. **Greene, E.A. and Spiegelman, G.B. (1996)** The Spo0A protein of *Bacillus subtilis* inhibits transcription of the *abrB* gene without preventing binding of the polymerase to the promoter. *J Biol Chem, 271, 11455-11461.*

49. **Rojo, F., Mencia, M., Monsalve, M. and Salas, M. (1998)** Transcription activation and repression by interaction of a regulator with the alpha subunit of RNA polymerase: the model of phage phi 29 protein p4. *Prog Nucleic Acid Res Mol Biol, 60, 29-46.*

50. **Jacob, F. and Monod, J. (1961)** Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol, 3, 318-356.*

51. **Crawford, I.P. and Stauffer, G.V. (1980)** Regulation of tryptophan biosynthesis. *Annu Rev Biochem, 49, 163-195.*

52. **Brennan, R.G. and Matthews, B.W. (1989)** The helix-turn-helix DNA binding motif. *J Biol Chem, 264, 1903-1906.*

53. **Winkler, W.C. and Breaker, R.R. (2005)** Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol, 59, 487-517.*

54. **Grundy, F.J., Yousef, M.R. and Henkin, T.M. (2005)** Monitoring uncharged tRNA during transcription of the *Bacillus subtilis glyQS* gene. *J Mol Biol, 346, 73-81.*

55. **Narberhaus, F., Waldminghaus, T. and Chowdhury, S. (2006)** RNA thermometers. *FEMS Microbiol Rev, 30, 3-16.*

56. **Masse, E., Majdalani, N. and Gottesman, S. (2003)** Regulatory roles for small RNAs in bacteria. *Curr Opin Microbiol, 6, 120-124.*

57. **Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J. and Huttenhofer, A. (2002)** Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus. Proc Natl Acad Sci U S A, 99, 7536-7541.*

58. **Tang, T.H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J.P. and Huttenhofer, A. (2005)** Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus. Mol Microbiol, 55, 469-481.*

59. **Winkler, W.C. (2005)** Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr Opin Chem Biol, 9, 594-602.*

60. **Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P. (2007)** CRISPR provides acquired resistance against viruses in prokaryotes. *Science, 315, 1709-1712.*

61. **Gottesman, S., McCullen, C.A., Guillier, M., Vanderpool, C.K., Majdalani, N., Benhammou, J., Thompson, K.M., FitzGerald, P.C., Sowa, N.A. and FitzGerald, D.J. (2006)** Small RNA regulators and the bacterial response to stress. *Cold Spring Harb Symp Quant Biol, 71, 1-11.*

62. **Mwangi, M.M. and Siggia, E.D. (2003)** Genome wide identification of regulatory motifs in *Bacillus subtilis*. *BMC Bioinformatics, 4, 18.*

63. **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 U S A, 99, 7323-7328.*

64. **Bussemaker, H.J., Li, H. and Siggia, E.D. (2001)** Regulatory element detection using correlation with expression. *Nat Genet, 27, 167-171.*

65. **Conlon, E.M., Liu, X.S., Lieb, J.D. and Liu, J.S. (2003)** Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A, 100, 3339-3344.*

66. **Keles, S., van der Laan, M. and Eisen, M.B. (2002)** Identification of regulatory elements using a feature selection method. *Bioinformatics, 18, 1167-1175.*
67. **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.*
68. **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.*
69. **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.*
70. **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.*
71. **Pavesi, G., Mauri, G. and Pesole, G. (2001)** An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics, 17 Suppl 1, S207-214.*
72. **van Helden, J., Andre, B. and Collado-Vides, J. (2000)** A web site for the computational analysis of yeast regulatory sequences. *Yeast, 16, 177-187.*
73. **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.*
74. **Lawrence, C.E. and Reilly, A.A. (1990)** An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins, 7, 41-51.*
75. **Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993)** Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science, 262, 208-214.*
76. **Neuwald, A.F., Liu, J.S. and Lawrence, C.E. (1995)** Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci, 4, 1618-1632.*
77. **Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A. et al. (2003)** Computational discovery of gene modules and regulatory networks. *Nat Biotechnol, 21, 1337-1342.*
78. **Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. and Palsson, B.O. (2004)** Integrating high-throughput and computational data elucidates bacterial networks. *Nature, 429, 92-96.*
79. **Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. (2002)** Combining location and expression data for principled discovery of genetic regulatory network models. *Pac Symp Biocomput, 437-449.*
80. **Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. et al. (2002)** Transcriptional regulatory networks in *Saccharomyces cerevisiae. Science, 298, 799-804.*
81. **Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004)** Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature, 431, 308-312.*
82. **Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002)** Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet, 31, 64-68.*
83. **Gutierrez-Rios, R.M., Rosenblueth, D.A., Loza, J.A., Huerta, A.M., Glasner, J.D., Blattner, F.R. and Collado-Vides, J. (2003)** Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res, 13, 2435-2443.*
84. **Herrgard, M.J., Covert, M.W. and Palsson, B.O. (2004)** Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol, 15, 70-77.*
85. **Newberg, L.A., Thompson, W.A., Conlan, S., Smith, T.M., McCue, L.A. and Lawrence, C.E. (2007)** A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics, 23, 1718-1727.*

CHAPTER 2

LacPlantReg: A database of *L. plantarum* WCFS1 regulatory proteins and context-based local regulons

Michiel Wels
Wendy Pluk
Michiel Kleerebezem
Roland J. Siezen

Acknowledgements

We would like to thank John Janssen for his help in predicting the conserved gene clusters between the genomes of *L. johnsonii*, *B. subtilis* and *L. plantarum*. We thank Christof Francke for his help in the analysis of the LacI family of transcriptional regulators.

Identification and analysis of the regulatory proteins is the first step in understanding the regulatory network underlying the adaptive mechanisms of a unicellular organism. The putative regulatory proteins of *Lactobacillus plantarum* were subjected to advanced annotation and divided in (sub)-families based on the data of different genome databases (COG, Pfam and ERGO). Regulons were predicted using genome context conservation and strain comparison data. The predicted regulons were grouped in classes based on biochemical function. Regulatory proteins within the same (sub)-family often appeared to regulate regulons containing genes with comparable molecular functions. Comparison with other lactic acid bacteria (LAB) genomes showed that *L. plantarum* has the highest fraction of regulatory proteins. The highest number of differences between regulatory proteins was observed in the families LacI, LysR and MarR. All three families can generally be correlated to adaptation to different or changing environments, i.e., regulation of energy metabolism and cell-envelope processes in relation to available nutrients (LacI and LysR) and regulation of protective mechanisms in response to harmful chemicals (MarR). The high number of regulatory proteins within these three regulatory families is in line with the flexibility and versatility of the species *L. plantarum*.

Introduction:

The capacity to adapt to environmental conditions is crucial for organisms to survive and prosper in different ecological niches. Highly versatile and flexible bacteria are known to contain a large repertoire of genetic potential that enables adaptive variations of the expressed biochemical pathways and cellular functions, allowing the organism to cope with the different environmental situations it encounters. Regulatory proteins control the expression of most genes and are in many cases activated by a specific signal (i.e. a metabolite or stress factor, etc.) and respond by modulation of their binding capacity to the DNA in the upstream region of the target genes, thereby affecting transcription of the downstream-located gene. Regulatory proteins are divided into different protein families on basis of the sequence similarity of the signaling and DNA-binding domains (1, 2). It has been shown that regulatory proteins within the same families tend to regulate the expression of genes with comparable functions. As an example, ArsR family regulators generally respond to high concentrations of different heavy metals (3), LacI family members to the availability of different sugar substrates (4) and MarR to harmful chemicals (5).

Regulation of gene expression in bacteria often occurs within a genetic locus, where both the regulatory protein and its target genes are genetically closely linked and encoded within a single genetic locus or operon. Analysis of the complete regulatory network of *Escherichia coli* showed that only very few (7 – 17) regulators were predicted to act as “global regulators” (6), while all other regulators were predicted to fulfill restricted and chromosomally localized regulatory functions. Nevertheless, the best-studied regulatory proteins in *Firmicutes* are known to act on a global level (7, 8), while only 8 regulators are known to regulate more than 20 genes in the best-studied *Firmicutes* species *Bacillus subtilis* (9). The other regulatory proteins control the expression of only a few genes, often organized in a single operon. These local regulators are often transcribed divergently from (10, 11), or as part of these operons (12). Therefore, analysis of (the level of conservation of) gene context of regulatory protein encoding genes should allow the accurate prediction of (part of) their regulon.

Within the lactic acid bacteria (LAB) *Lactobacillus plantarum* is a renowned flexible and versatile species that is encountered in a variety of different environmental niches. This ecological flexibility of *L. plantarum* is reflected by its genome size that is among the largest known in LAB (13). To gain insight in the adaptive abilities of *L. plantarum*, the underlying mechanisms of adaptation need to be studied. Adaptation is largely regulated at the level of gene transcription and regulatory proteins play a key role in these regulatory processes. The sequencing of the genome of *L. plantarum* WCFS1 (14) and various other LAB in recent years (15, 16) enabled the comparative analysis of the full complement of regulatory proteins in *L. plantarum* in relation to other LAB.

Previous studies have shown that the relation between genome size and the number of regulatory proteins does not grow in a linear fashion, as is the case for other general functions of a microbe like metabolic enzymes or transporters, but rather in an exponential way. Consequently, with increasing genome sizes, a larger portion of the genome becomes devoted to the regulation of transcription (17). In addition it was shown that flexibility of the organism (free-living, pathogenic or endosymbiont), although being linked to genome size, shows to be an additional determinant for the number of regulatory proteins an organism possesses (18). In this study we show that the number of regulatory proteins in LAB genomes is also related to genome size and niche properties. Furthermore, we show that the difference in the number of regulators is particularly apparent for specific families of regulatory proteins, which appears to be related to the functions of the genes that are commonly regulated by these regulator families.

Methods:

A flow scheme describing the followed procedure is shown in Figure 1. Tools used at the specific steps are indicated in the figure.

Selection of regulatory proteins

Regulatory proteins were gathered from the original annotation of the *L. plantarum* WCFS1 genome (14). A protein was classified as a regulatory protein if it met one of the following criteria; 1) member of the main class “Regulatory Functions”; 2) contained the words “regulator” or “regulation” in the function description; 3) member of the sub class “Transcription factors” within the main class “Transcription”. This resulted in the identification of 286 putative regulatory proteins. In addition, three *L. plantarum* proteins annotated as “regulatory protein” in the ERGO bioinformatics suite (19), but missed in the original annotation, were added leading to a total set of 289 regulatory proteins.

Advanced annotation of regulatory proteins

To determine the characteristics of the *L. plantarum* regulatory proteins, a comparison was made with the genome of *B. subtilis*, for which most experimental data is available regarding its regulatory network (20). Bidirectional best hits (BBH) between the genomes of *L. plantarum* and *B. subtilis* were determined using BLAST (21). Gene names and PFAM domains (1) with an e-value below $1e^{-10}$ were extracted from a PEDANT automatic annotation (22). COG (Clusters of Orthologous Groups) were predicted using

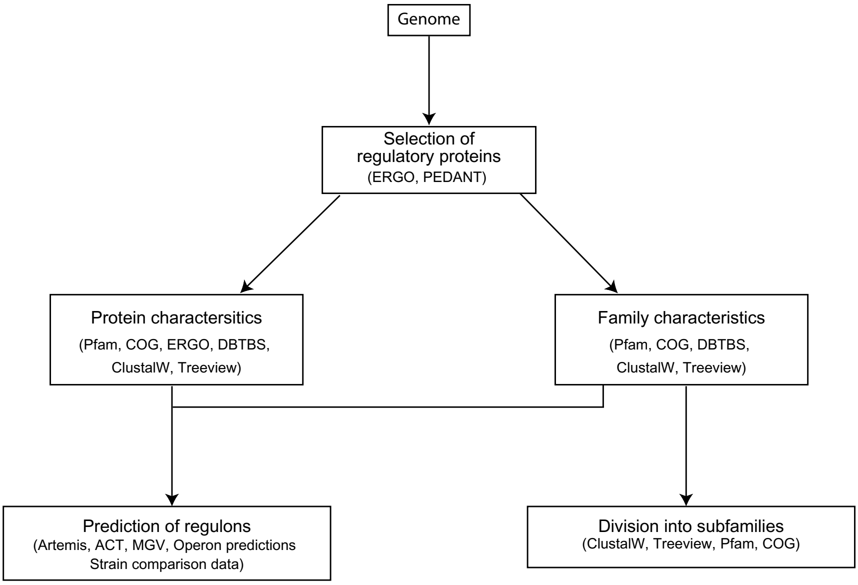

Genome
Selection of regulatory proteins (ERGO, PEDANT)
Protein charactersitics (Pfam, COG, ERGO, DBTBS, ClustalW, Treeview)
Family characteristics (Pfam, COG, DBTBS, ClustalW, Treeview)
Prediction of regulons (Artemis, ACT, MGV, Operon predictions Strain comparison data)
Division into subfamilies (ClustalW, Treeview, Pfam, COG)

Figure 1: Flow chart describing the procedure for in-depth annotation of the regulatory proteins.

BLAST against the COG database (2). When the ten best hits of the protein against the COG database were part of the same COG, the protein was classified in that COG. In some cases, multiple COGs were assigned to different regions of one protein to annotate fusion proteins. Additional information regarding the function, classification or name of the genes found by ERGO, BBH, PFAM or COG analysis, were added to the annotation information of the protein in the *L. plantarum* WCFS1 database.

Regulatory protein (sub) family division and analysis

Regulatory proteins were divided into different families on basis of the original annotation in combination with additional information derived from ERGO, PFAM or COG analysis. COG and PFAM domains were considered family-specific if the domain was assigned to more than half of the proteins within a family. Protein sequences of the different families were aligned using ClustalW (23) and used for construction of bootstrap neighbor-joining trees. Trees were converted to image files to allow visualization in a web-browser. Known protein families were divided into subfamilies on basis of sequence alignment, bootstrap tree, PFAM, COG and/or sequence length. When patterns were observed within a family (for example a conserved absence/presence pattern of two different COGs), the proteins were divided into different subfamilies.

Prediction and classification of regulons

Conserved gene context (clusters) was predicted for *L. plantarum* on basis of shared genome context with the genome sequences of *B. subtilis subsp. subtilis* 168 (24) and *L. johnsonii* NCC 533 (25). All clusters containing at least one regulatory protein were manually inspected and curated using Artemis (26) and the Microbial Genome Viewer (27). To increase cluster prediction reliability, comparative genome hybridization data related to the genomic diversity among *L. plantarum* strains (28) was used to identify conserved presence/absence patterns between regulators and proposed regulons. Candidate global regulators and their regulons were predicted from published experimental data using the gene name of the bidirectional best hit in *B. subtilis* as the literature query input. The predicted regulons were added to the information for the regulatory protein in the *L. plantarum* WCFS1 database.

The functional classification available in the *L. plantarum* WCFS1 database (14) was used to functionally classify the predicted regulons. Regulons were appointed to functional classes, when genes in a regulon were found to belong to a single and coherent functional class. Alternatively, regulons encoding multiple classes were classified as "Mixed function". The presence of genes classified as "Transport and binding proteins", "Hypothetical proteins" or "Regulatory functions" was not primarily used for regulon functional classification, unless a regulon consisted of only "Hypothetical proteins" and/or "Transport and binding proteins". In addition, no classification was assigned when regulons contained only proteins classified as "Regulatory functions".

Comparison of L. plantarum regulators with other lactic acid bacteria

Data describing the orthologous relations (LaCOG and COG annotation) of the genes within the complete genomes of 12 lactic acid bacteria (*L. plantarum* WCFS1, *L. johnsonii* NCC 533, *Oenococcus oeni* PSU-1 (29), *Lactococcus lactis* IL-1403, *Lactococcus lactis subsp. cremoris* SK11, *Streptococcus thermophilus* LMD-9 *Lactobacillus brevis* ATCC 367, *Lactobacillus casei* ATCC 334, *Lactobacillus gasseri* ATCC 33323, *Leuconostoc mesenteroides subsp. mesenteroides* ATCC 8293, *Pediococcus pentosaceus* ATCC 25745, *Lactobacillus delbrueckii subsp. bulgaricus* BAA-365 (all (16)), was obtained from Makarova et al (30).

Results:

L. plantarum regulators and regulons

The putative regulatory proteins of *L. plantarum* were selected on basis of the original genome annotation of *L. plantarum*, in combination with up-to-date annotation databases ERGO (19) and PEDANT (31). COGs were chosen over PFAM to classify the regulatory proteins as they cover larger parts of the protein. The families were regrouped in subfamilies if multiple proteins (>2) shared the same COG. If possible, a prediction of the name or type of family was made based on annotation information and shared COGs/PFAM domains. Twenty-eight proteins of the original category 'other' were grouped into four new (sub)-families (PadR, HipB, RRF2 and a second MarR family). The families DeoR, GntR, Two-component regulators and Cell division were also divided into subfamilies. An overview of the different (sub-) families and the number of proteins grouped into these families is shown in Table 1. The full list of proteins and details on their annotation can be found at http://www.cmbi.ru.nl/regulator_database .

On basis of their genome context, putatively coregulated genes (regulon) were predicted for all regulatory proteins (see Material and Methods for details). A size distribution of the predicted local regulons (Figure 2) revealed that almost half of the regulons (49%) regulate only 2-3 genes, while regulatory proteins that control the expression of more than 9 genes are rare (2% of the total set). For only 8% of the regulators no regulon could be predicted (size=1 in Figure 2). Regulons were grouped on basis of encoded biochemical functions (Figure 3). If genes belonged to multiple classes, the regulons were assigned "mixed function": A total of 15% of the regulons

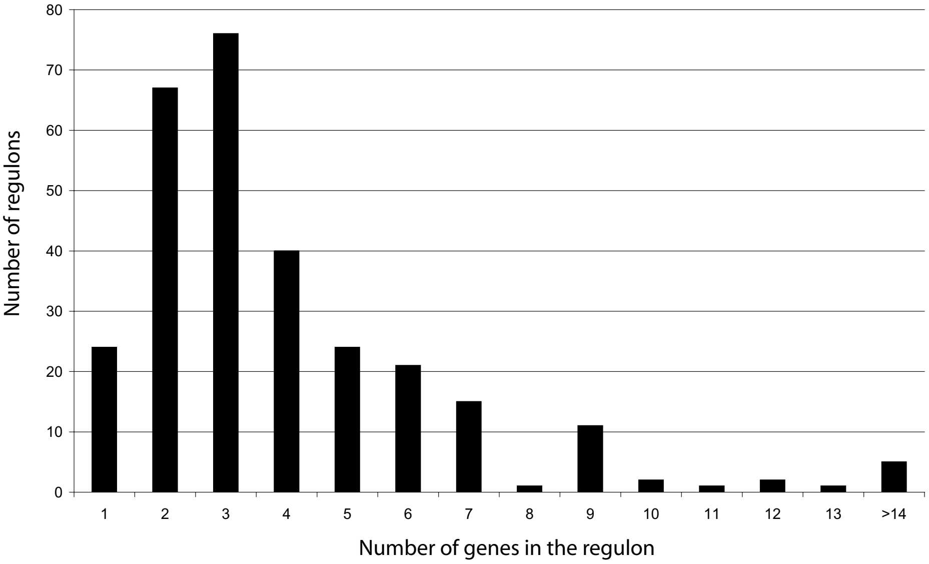


Figure 2: Size distribution of the different regulons.
Regulons in this plot consist of maximally two different Transcriptional Units.

were part of the "mixed function" category, while 23% were part of the "hypothetical proteins" category. Other major categories that appeared to be represented in dedicated regulons were energy metabolism (19%), and cell envelope (6%). For 16% of the regulons, no function could be assigned. An overview of all predicted regulons can be found at http://www.cmbi.ru.nl/regulator_database.

LysR

One of the largest families of transcriptional regulators found in *L. plantarum* is LysR (22 members; table 1). Of these 22 regulons, 5 are involved in regulation of regulons classified in energy metabolism, one in cell-envelope processes, one in amino acid metabolism and eight regulons were classified as hypothetical. For the other regulatory proteins it was impossible to define a clear class due to the lack of genome context (3 cases) or a too diverse make-up of the functions of the regulated genes (mixed: 4 cases). The high number of regulatory proteins involved in regulation of energy metabolism genes is in good agreement with literature. For example, in *B. subtilis* LysR-type regulators were associated with regulation of alpha-acetolactate synthesis (AlsR (10)) and the TCA cycle (CitR (32) and CcpC (33)). Other studies suggest that LysR-type regulators control the expression of amino acid biosynthesis genes (34). In *L. plantarum* we observed one of the LysR regulatory proteins (lp_3495) to be involved in the regulation of aromatic amino acid biosynthesis genes (*aroC2* and *aroD2*).

MarR

MarR regulators (5) were also found in high numbers (21 members) in *L. plantarum*. Members of the MarR family are frequently found to be involved in resistance to a variety of different harmful chemicals, including antibiotics, organic solvents and oxidative agents and radicals (35). MarR-type regulators

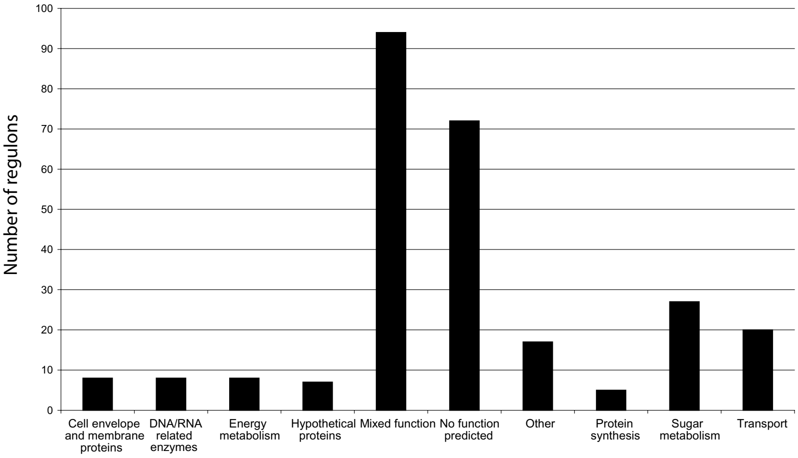


Figure 3: Functional classification of all regulons based on shared biochemical functions.
For 45 regulators no classification could be made.

have also been identified in the regulatory circuits underlying virulence control in various species, including *Salmonella typhimurium* (SlyA (36)), *Vibrio cholerae* (RovA (37)), several *Erwinia* species (PecS and Hor (38, 39)) and *Staphylococcus aureus* (MgrA (40)). Seven of the MarR regulons in *L. plantarum* could not be functionally classified, since they consisted solely of hypothetical proteins or only the regulator. Five MarR-type regulons were found to be involved in cell-envelope processes, which could relate to a response mechanism to compounds that damage the cell envelope. Furthermore, of these five regulons, three are involved in synthesis of different cell-envelope components: a regulon consisting of *murB* (*lp_0814*) coding for a UDP-N-acetylmuramate dehydrogenase (peptidoglycan biosynthesis), one consisting of *licD* (*lp_0844*) coding for a lipopolysaccharide biosynthesis protein and a regulon consisting of the gene cluster *lp_1816* - *lp_1819*, coding for several different teichoic acid biosynthesis proteins. Interestingly, these three different regulons encode genes that play a role in the biosynthesis of different components of the cell envelope and all seem to be regulated by a different member of the MarR family. In addition, we found 3 regulons classified as involved in transport, of which 2 encode multidrug-resistance transporters (MDR). Next to these 2 regulons we also observed the presence of another MDR in a regulon classified as "Energy Metabolism" on basis of the presence of a poorly classified nitroreductase. These three regulons are probably involved in the extrusion of damaging chemicals. Finally, one MarR-type regulon was classified as fatty acid metabolism, which could also relate to chemical damage response mechanisms by controlling cell membrane damage repair. However, the gene in this regulon (*lp_0955*) is only generally annotated as an acyl-carrier protein phosphodiesterase. Therefore the link between this gene and cell-membrane repair remains to be established. Of the remaining four regulons, two were additionally classified as "Energy metabolism", and the remaining two as "DNA metabolism" and "Cellular processes". These latter two regulons could potentially be involved in responses to damaging agents that have entered the cell; the first regulon encodes an A/G-specific adenine glycosylase (*lp_3349*, *mutY*), involved in DNA recombination and repair and the second regulon contains a gene encoding a small heat-shock protein (*lp_0129*, *hsp1*).

LacI

Another interesting class of regulators is the LacI-family (41). In fact, the master regulator of catabolite repression in gram-positive bacteria, CcpA, is a well-known member of the LacI family (42).

Despite the fact that CcpA is a renowned global regulator, the other LacI-family members, for instance those that have been characterized in *B. subtilis*, appear to act rather locally (43). LacI-family members were shown previously to regulate the expression of operons related to sugar metabolism. Indeed, six out of the fifteen LacI-type regulons found in *L. plantarum* were related to energy metabolism, in many cases in the subclass sugars. These regulons consist mainly of genes encoding transport functions like PTSs that allow active and specific internalization of the available sugar. Of the other regulons predicted to be under LacI-type regulator control, four were assigned to mixed function, two to central intermediary metabolism, and one to hypothetical proteins and protein fate (*pepQ*). The latter gene seems to be unrelated to sugar metabolism. However, the predicted regulator of this regulon is CcpA that is well-known to control the expression of a large number of genes with very diverse functions (44). Moreover, regulation of the expression of *pepQ* by CcpA was shown in *L. lactis* (41), *L. pentosus* (45) and *L. delbrueckii* (46). Finally, only for a single LacI-type regulator, no corresponding regulon could be predicted.

The COG annotation of the proteins of *L. plantarum* and 11 other lactic acid bacteria were gathered from Makarova et al (30). On basis of the COG annotation, the proteins were divided into the (sub)-families predicted by our analysis (Table 1). All (sub)-families lacking a coherent COG annotation in our analysis were not analyzed. Sixty differences were observed for *L. plantarum* between our prediction and Makarova et al. Despite this large overall difference, 11 out of 26 (sub-) families were common between the two predictions, whereas in 5 only one difference was observed. Fourteen of the proteins found by Makarova et al in addition to our curated prediction are annotated as “hypothetical proteins”. Detection of these is probably the result of an increase in knowledge of the function of these proteins since the original annotation in 2003. In other cases, it can be doubted if the protein is a regulatory protein. As an example eighteen additional members of COG1396 (HipB family) are predicted by Makarova et al., for which eight are annotated as prophage proteins and not likely to function in regulation. Another COG that had a great difference between the Makarova annotation and our set was COG2207 (AraC family). This COG has only one member in the Makarova dataset, while 8 were predicted with our analysis. Of the 7 additional members found only by our prediction, 6 were already annotated as AraC-type regulators in the original annotation (14). Further analysis showed that all these additional members were part of COG4753, a COG not present in the COG annotation used for the manually curated dataset. The Makarova et al. data was based on a newer COG annotation (47) and did contain COG4753.

Comparative analysis of regulators in LAB

When comparing *L. plantarum* with the other LAB (table 1), it is clear that *L. plantarum* has the largest set of regulatory proteins, both in absolute numbers (230 according to Makarova et al., 206 in our method) as well as in percentage of the total number of predicted

Table 1: Regulatory protein distribution in 12 lactic acid bacteria genomes based on data from Makarova et al. (30). Lacpl; L. plantarum WCFS1 (Between brackets are the regulatory proteins assigned in our analysis), Lacbr; L. brevis ATCC 367, Lacca; L. casei ATCC 334, Lacla; Lactococcus lactis IL-1403, Laccr; Lactococcus lactis subsp. cremoris SK11, Lacjo; L. johnsonii NCC533, Lacga; L. gasseri ATCC 33323, Leume; Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293, Pedpe; Pediococcus pentosaceus ATCC 25745, Oenoe; Oenococcus oeni PSU-1, Lacde; L. delbrueckki subsp. bulgaricus BAA-365, Strth; Streptococcus thermophilus LMD-9. Species are ordered on basis of the percentage of genes encoding regulatory proteins compared to the total number of genes.

*Proteins that are part of a second MarR family.

Family	subfamily	COG	Lacpl	Lacbr	Lacjo	Pedpe
2-component regulators	(BaeS)	COG0642	5 (8)	7	5	4
2-component regulators	(OmpR)	COG0745	7 (6)	7	6	5
2-component regulators	(LytT/AlgR)	COG3279	8 (4)	3	7	3
AraC	-	COG2207	1 (8)	3	0	0
ArsR	-	COG0640	7 (9)	6	3	2
BglB	-	COG3711	10 (5)	0	4	2
Cell division	-	COG0537	3 (2)	2	2	2
Cell division	-	COG2337	2 (1)	1	1	1
Crp/FNR	-	COG0664	3 (3)	4	1	3
DeoR	-	COG2390	4 (4)	2	1	1
DeoR	(GlpR)	COG1349	5 (5)	0	2	1
GntR	-	COG1725	4 (4)	3	1	2
GntR	(PhnF)	COG2188	10 (10)	4	8	10
LacI	-	COG1609	15 (15)	10	7	7
LysR	-	COG0583	21 (22)	7	6	8
LytR	-	COG1316	4 (4)	3	4	2
MarR	-	COG1846	21 (21)	22	5	7
MerR	-	COG0789	14 (10)	12	4	5
Other	(PadR)	COG1695	4 (4)	4	0	2
Other	(MarR*)	COG1733	7 (7)	5	1	3
Other	(HipB)	COG1396	24 (8)	13	20	8
Other	(RRF2)	COG1959	4 (4)	5	1	1
protein interactions	-	COG1393	5 (5)	3	2	2
RpiR	-	COG1737	7 (7)	2	6	7
Sugars	-	COG1940	9 (5)	7	5	7
TetR-ArcR	-	COG1309	26 (25)	19	6	6
Total			230 (206)	154	108	101
Genome size (nt)			3308274	2291220	1992676	1832387
Total number of genes			3007	2185	1821	1755
Percentage of complete genome			7.6 %	7.1 %	5.9 %	5.8 %

Lacga	**Lacla**	**Lacca**	**Laccr**	**Oenoe**	**Leume**	**Lacde**	**Strth**
4	5	10	6	4	6	5	2
5	6	12	8	5	7	6	4
4	1	3	2	2	3	3	4
0	1	0	1	2	2	2	0
2	2	5	2	2	2	1	0
4	3	12	3	2	2	0	0
1	2	3	2	2	2	1	1
1	0	1	0	1	1	0	0
2	2	1	2	1	2	2	1
1	1	1	0	1	0	2	0
3	3	5	4	0	0	2	1
1	0	4	0	1	1	2	1
9	2	6	3	4	3	2	1
5	7	8	6	8	13	2	3
7	7	8	5	7	5	8	3
3	6	3	3	2	4	4	3
6	12	8	9	11	8	3	4
5	5	6	8	5	3	3	2
0	3	5	3	1	2	0	0
2	2	2	4	1	3	1	0
16	26	21	19	7	16	6	10
0	1	0	2	1	2	2	1
2	7	2	7	0	1	2	3
5	8	6	7	2	1	1	0
3	3	2	4	2	3	3	3
7	14	16	14	13	9	6	3
98	129	150	124	87	101	69	50
1894360	2365589	2895264	2438589	1780517	2038396	1856951	1856368
1755	2321	2751	2384	1691	1970	1721	1710
5.6 %	5.6 %	5.5 %	5.2 %	5.1 %	5.1 %	4.0 %	2.9 %

genes on the genome (7.64%). Previous studies on the relation between genome size and the number of regulators already showed that not only absolute numbers, but also the fraction of regulatory proteins increases with genome size; this is in clear contrast to several other gene families like transporters and small molecule metabolism (17, 18). Since *L. plantarum* harbors the largest genome known among the LAB used in this study, our observations are in line with this previous study. Next to the relation of genome size to the percentage of regulators, a second difference between the genomes is observed. On basis of the percentage of genes dedicated to regulatory proteins, three distinct classes can be formed. Most LAB (8 species) have approximately the same percentage (5.1 – 5.9 %) of their genes dedicated to regulation. However, *L. plantarum* and *Lactobacillus brevis* have a distinctly higher percentage of their genome dedicated to regulation (7.6% and 7.1% respectively) while on the contrary, *Streptococcus thermophilus* and *Lactobacillus delbrueckii* have a far lower than average percentage of their genome assigned to regulation (2.9% and 4.0% respectively). This division seems to be linked to the lifestyle or niche of the different species. *L. delbrueckii* and *S. thermophilus* are well-known to be extremely dedicated to their relatively stable and nutrient-rich environment (milk), whereas *L. plantarum* and *L. brevis* are among the most flexible LAB species known, are highly adaptive, and can be found in different niches. This relation between the fraction of regulatory proteins and lifestyle or niche has been previously described for a set of organisms that included endosymbionts, pathogens, extremophiles and free-living organisms and displayed the same relation; the more diverse the lifestyle of an organism, the larger number of regulatory proteins it possesses (18). Although the increase in genome size was responsible for part of the link between lifestyle and fraction of regulatory proteins (free-living organisms also tend to have a larger genome), the authors showed that lifestyle on itself is also a determinant factor.

The difference in the numbers of regulatory proteins is observed in all families, but in different proportions. The biggest difference is observed for the MarR family: 21 and 22 members in *L. plantarum* and *L. brevis* versus maximally 9 for the other LAB. Other differences are observed within the regulatory families of LysR, MerR, LacI and TetR-ArcR. LysR is particularly large in *L. plantarum* (22 members) while the differences in LacI and TetR are more spread among the species. The AraC family and the families related to cell division seem to be rather constant in size.

Conclusions and Discussion:

A database was constructed of all predicted regulatory proteins encoded in the genome sequence of *L. plantarum* in combination with part of their predicted local regulons. The identified regulatory proteins were annotated in more detail using PFAM, COG and the ERGO bioinformatics suite. The collective information gathered for each regulator protein was used to designate it within a regulator family and/or subfamily.

The family prediction of this in-depth annotation was compared with the data from Makarova et al. (30). This comparison resulted in many differences within the protein families, several of which could be assigned to technical differences based on the use of different COG database releases (e.g. the difference within the AraC family). However, in other cases the inconsistencies are on a different level. When comparing the BglB family prediction of both methods, Makarova et al. predict ten proteins to be member of this regulatory family, while our annotation predicts only five members to be part of the family. Three out of the five additional genes found by Makarova et al *(lp_3543, lp_3656 and lp_3622)* encode assigned regulators but were not assigned to the family specific COG (COG3711). The COG was assigned to the other two genes, but was not incorporated in this family on basis of other criteria. For *lp_3138*, an additional COG was assigned (COG1762) specific for a EIIA domain of a PTS system. The other gene (*lp_0231*) was annotated to be part of a different regulatory family (DeoR) by ERGO. Therefore, all these five proteins were regarded as Regulatory proteins, but included into the "Other" class, as different resources suggested different family division.

When comparing the number of regulatory proteins in *L. plantarum* with other LAB, major differences were observed in three families (MarR, LysR and LacI). For all three of these families, the relation with adaptation to the environment is evident; LysR and LacI regulate genes that are involved in energy metabolism. A broader capacity to metabolize different compounds is directly related to survival in variable environments. Consistently, LAB that are known to occur in only few, and generally very stable environments like milk (*S. thermophilus, L. delbrueckii*) have very few LacI and LysR-type regulators. Differences encountered for the MarR family are probably indicative for more subtle environmental adaptation processes. A broader ecological adaptation capacity leads to increased variation of the harmful chemicals encountered. Therefore it seems plausible that more complex environments like the gastrointestinal tract and plant surfaces require a larger number of genes that protect against these chemicals. As these genes are regulated by specific classes of regulatory proteins, their number increases as well. If an organism is adapted to a more stable environment, like milk, it will gradually lose these protection systems and with that the regulatory proteins responsible for sensing the damaging agent.

Local regulons were predicted on basis of conserved gene order with the complete genomes of *L. johnsonii* and *B. subtilis*. On basis of the genome context, regulons were predicted for 84% of the regulatory proteins. Although these regulons can be incomplete, classification provides insight in the processes controlled by the regulator. For 45 regulators we observe that genome context was not conserved and could therefore not be used for regulon prediction. In these cases it can be argued whether these regulatory proteins even act on a local level or if they control the expression of more distantly located genes. The same question can of course also be addressed for the cases where we observe conserved gene context. However, the observed correlation in function between the regulons derived from literature and the genome context predictions from this study suggests that in many cases these regulators seem to act on (at least) a local level. To investigate whether a more global regulation is also observed for these regulators, localization of the regulatory binding site would be of great assistance. Analysis of the upstream regions of the locally controlled operons could act as a good starting point for finding the regulatory binding site of this transcriptional regulator. On basis of the size of the regulons (generally only 2-3 proteins, which are often in the same operon), comparative genomics methods like phylogenetic footprinting (48, 49) would be the best strategy for finding the consensus binding sites for these regulatory proteins. Ultimately, regulon and corresponding regulator protein predictions can also be validated by (meta) analyses of transcriptomics data.

All data was combined into a database (LacPlantReg) that allows querying of the data in a user-friendly interface. The database can be accessed at http://www.cmbi.ru.nl/regulator_database.

References

1. **Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. and Durbin, R. (1998)** Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res, 26, 320-322.*
2. **Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997)** A genomic perspective on protein families. *Science, 278, 631-637.*
3. **Busenlehner, L.S., Pennella, M.A. and Giedroc, D.P. (2003)** The SmtB/ArsR family of metalloregulatory transcriptional repressors: Structural insights into prokaryotic metal resistance. *FEMS Microbiol Rev, 27, 131-143.*
4. **Nguyen, C.C. and Saier, M.H., Jr. (1995)** Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. *FEBS Lett, 377, 98-102.*
5. **Wilkinson, S.P. and Grove, A. (2006)** Ligand-responsive transcriptional regulation by members of the MarR family of winged helix proteins. *Curr Issues Mol Biol, 8, 51-62.*
6. **Martinez-Antonio, A. and Collado-Vides, J. (2003)** Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol, 6, 482-489.*
7. **Miwa, Y., Nakata, A., Ogiwara, A., Yamamoto, M. and Fujita, Y. (2000)** Evaluation and characterization of catabolite-responsive elements (cre) of *Bacillus subtilis*. *Nucleic Acids Res, 28, 1206-1210.*
8. **Molle, V., Nakaura, Y., Shivers, R.P., Yamaguchi, H., Losick, R., Fujita, Y. and Sonenshein, A.L. (2003)** Additional targets of the *Bacillus subtilis* global regulator CodY identified by chromatin immunoprecipitation and genome-wide transcript analysis. *J Bacteriol, 185, 1911-1922.*
9. **Lozada-Chavez, I., Janga, S.C. and Collado-Vides, J. (2006)** Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res, 34, 3434-3445.*
10. **Renna, M.C., Najimudin, N., Winik, L.R. and Zahler, S.A. (1993)** Regulation of the *Bacillus subtilis alsS, alsD*, and *alsR* genes involved in post-exponential-phase production of acetoin. *J Bacteriol, 175, 3863-3875.*
11. **Sun, D. and Setlow, P. (1993)** Cloning and nucleotide sequence of the *Bacillus subtilis ansR* gene, which encodes a repressor of the ans operon coding for L-asparaginase and L-aspartase. *J Bacteriol, 175, 2501-2506.*
12. **Sato, T. and Kobayashi, Y. (1998)** The *ars* operon in the skin element of *Bacillus subtilis* confers resistance to arsenate and arsenite. *J Bacteriol, 180, 1655-1661.*
13. **Chevallier, B., Hubert, J.C. and Kammerer, B. (1994)** Determination of chromosome size and number of rrn loci in Lactobacillus plantarum by pulsed-field gel electrophoresis. *FEMS Microbiol Lett, 120, 51-56.*
14. **Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W. et al. (2003)** Complete genome sequence of Lactobacillus plantarum WCFS1. *Proc Natl Acad Sci U S A, 100, 1990-1995.*
15. **Liu, M., van Enckevort, F.H. and Siezen, R.J. (2005)** Genome update: lactic acid bacteria genome sequencing is booming. *Microbiology, 151, 3811-3814.*
16. **Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N. et al. (2006)** Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A, 103, 15611-15616.*
17. **van Nimwegen, E. (2003)** Scaling laws in the functional content of genomes. *Trends Genet, 19, 479-484.*
18. **Cases, I., de Lorenzo, V. and Ouzounis, C.A. (2003)** Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol, 11, 248-253.*

19. **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.*
20. **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, D75-77.*
21. **Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990)** Basic local alignment search tool. *J Mol Biol, 215, 403-410.*
22. **Riley, M.L., Schmidt, T., Artamonova, II, Wagner, C., Volz, A., Heumann, K., Mewes, H.W. and Frishman, D. (2007)** PEDANT genome database: 10 years online. *Nucleic Acids Res, 35, D354-357.*
23. **Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994)** CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res, 22, 4673-4680.*
24. **Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S. et al. (1997)** The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature, 390, 249-256.*
25. **Pridmore, R.D., Berger, B., Desiere, F., Vilanova, D., Barretto, C., Pittet, A.C., Zwahlen, M.C., Rouvet, M., Altermann, E., Barrangou, R. et al. (2004)** The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. *Proc Natl Acad Sci U S A, 101, 2512-2517.*
26. **Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B. (2000)** Artemis: sequence visualization and annotation. *Bioinformatics, 16, 944-945.*
27. **Kerkhoven, R., van Enckevort, F.H., Boekhorst, J., Molenaar, D. and Siezen, R.J. (2004)** Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics, 20, 1812-1814.*
28. **Molenaar, D., Bringel, F., Schuren, F.H., de Vos, W.M., Siezen, R.J. and Kleerebezem, M. (2005)** Exploring *Lactobacillus plantarum* genome diversity by using microarrays. *J Bacteriol, 187, 6119-6127.*
29. **Mills, D.A., Rawsthorne, H., Parker, C., Tamir, D. and Makarova, K. (2005)** Genomic analysis of Oenococcus oeni PSU-1 and its relevance to winemaking. *FEMS Microbiol Rev, 29, 465-475.*
30. **Makarova, K.S. and Koonin, E.V. (2006)** Evolutionary genomics of lactic acid bacteria. *J Bacteriol. 189, 1199 - 1208*
31. **Frishman, D., Mokrejs, M., Kosykh, D., Kastenmuller, G., Kolesov, G., Zubrzycki, I., Gruber, C., Geier, B., Kaps, A., Albermann, K. et al. (2003)** *The PEDANT genome database. Nucleic Acids Res, 31, 207-211.*
32. **Jin, S. and Sonenshein, A.L. (1994)** Transcriptional regulation of *Bacillus subtilis* citrate synthase genes. *J Bacteriol, 176, 4680-4690*.
33. **Libby, S.J., Mani, N., Bowe, F. and Sonenshein, A.L. (2000)** A cytolysin encoded by *Salmonella* is required for survival within macrophages. *J Mol Biol, 91, 865-878.*
34. **Henikoff, S., Haughn, G.W., Calvo, J.M. and Wallace, J.C. (1988)** A large family of bacterial activator proteins. *Proc Natl Acad Sci U S A, 85, 6602-6606.*
35. **Alekshun, M.N. and Levy, S.B. (1999)** The *mar* regulon: multiple resistance to antibiotics and other toxic chemicals. *Trends Microbiol, 7, 410-413.*
36. **Libby, S.J., Goebel, W., Ludwig, A., Buchmeier, N., Bowe, F., Fang, F.C., Guiney, D.G., Songer, J.G. and Heffron, F. (1994)** A cytolysin encoded by *Salmonella* is required for survival within macrophages. *Proc Natl Acad Sci U S A, 91, 489-493.*
37. **Revell, P.A. and Miller, V.L. (2000)** A chromosomally encoded regulator is required for expression of the *Yersinia enterocolitica inv* gene and for virulence. *Mol Microbiol, 35, 677-685.*
38. **Reverchon, S., Rouanet, C., Expert, D. and Nasser, W. (2002)** Characterization of indigoidine biosynthetic genes in Erwinia chrysanthemi and role of this blue pigment in pathogenicity. *J Bacteriol, 184, 654-665.*
39. **Thomson, N.R., Cox, A., Bycroft, B.W., Stewart, G.S., Williams, P. and Salmond, G.P. (1997)** The *rap* and *hor* proteins of *Erwinia*, *Serratia* and *Yersinia*: a novel subgroup in a growing superfamily of proteins regulating diverse physiological processes in bacterial pathogens. *Mol Microbiol, 26, 531-544.*
40. **Ingavale, S. (2005)** CcpA-dependent carbon catabolite repression in bacteria. *Infect Immun, 73, 475-490.*
41. **Zomer, A.L. and Saier, M.H., Jr. (1995)** Time-resolved determination of the CcpA regulon of

bacterial transcription factors. *FEBS Lett, 377, 98-102.*

42. **Deutscher, J., Francke, C. and Postma, P.W. (2006)** How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiol Mol Biol Rev, 70, 939-1031.*
43. **Mahr, K., Hillen, W. and Titgemeyer, F. (1998)** Carbon catabolite repression in *Bacillus subtilis. J Bacteriol, 66, 491-497.*
44. **Warner, J.B. and Lolkema, J.S. (2003)** CcpA-dependent carbon catabolite repression in bacteria. *Microbiol Mol Biol Rev, 67, 475-490.*
45. **Mahr, K., Hillen, W. and Titgemeyer, F. (2000)** Carbon catabolite repression in *Lactobacillus pentosus*: analysis of the ccpA region. *Appl Environ Microbiol, 66, 277-283.*
46. **Schick, J., Weber, B., Klein, J.R. and Henrich, B. (1999)** PepR1, a CcpA-like transcription regulator of *Lactobacillus delbrueckii subsp. lactis. Microbiology, 145 (Pt 11), 3147-3154.*
47. **Cases, I., de Lorenzo, V. and Tatusova, T.A. (2001)** The COG database: new developments in bacteria. *Trends Microbiol, 29, 22-28.*
48. **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.*
49. **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.*

CHAPTER 3

In silico promoter analysis in the intergenic regions of *L. plantarum* WCFS1

Michiel Wels
Michiel Kleerebezem
Roland J. Siezen

Manuscript in preparation

Conserved intergenic elements were predicted by comparative sequence analysis of all upstream regions of the *L. plantarum* WCFS1 genome. From the ten best-conserved motifs, nine were similar to DNA-protein binding sites known in other bacteria. Seven of these known motifs, as well as the only newly identified motif described here, are involved in processes related to transcription or translation. The residual three motifs appeared to be part of the inverted repeats flanking the previously identified transposons ISP1 and ISP2. Co-occurrence of more than one motif in a single upstream region predominantly resulted from different representations of the canonical sigmaA (SigA)-dependent promoter, and from separate identification of different parts of the T-box regulatory element. The promoter elements identified were used to predict the full complement of SigA promoters present in the genome of *L. plantarum*. Using the canonical spatial preference of these promoters and their relative distance to the start codon of the downstream-located gene, promoters could be predicted with a significance level of 85%. Applying this significance cut-off, allowed identification of sigA promoters in the upstream regions of 45% of the predicted transcriptional units.

Introduction:

Prokaryotic genomes are known to have a limited percentage of non-coding or intergenic DNA (1). These intergenic DNA regions are known to contain elements involved in transcription or translation processes of the flanking (coding) regions. Many of these elements, such as promoters (sigma factor binding sites) (2) or Shine-Dalgarno (SD) sequences (ribosome-binding sites) (3) are essential for binding of the mRNA or protein synthesis machinery, respectively. The requirement of these elements is illustrated by the use of the SD sequence in gene prediction software in order to accurately predict gene starts (4). Other intergenic elements, known as *cis*-regulatory binding sites can, alter the level of transcription or translation of the downstream gene by the activity of a regulating *trans*-acting factor.

Studies to detect conserved intergenic elements have been performed in the past and frequently employed consensus sequences based on experimentally determined binding sites (5,6). As only few of the intergenic *cis*-acting elements present in *L. plantarum* are based on experimental verification, comparative genomics or other *ab inito* methods will be necessary to predict intergenic conserved elements in this species. Although comparative genomic approaches have proven successful in finding *cis*-regulatory elements (7-9), they do not seem well-suited for the prediction of global regulators or other highly abundant *cis*-elements. Such elements are probably better characterized by analysis of the full complement of intergenic regions in one species.

In this study a large scale and computational intensive analysis of the intergenic regions of *L. plantarum* was performed by studying the upstream regions of all predicted genes. This analysis resulted in the identification of the ten most abundantly conserved intergenic sequence elements in the *L. plantarum* genome. Not surprisingly, the best conserved elements have sequence similarity to well-known and conserved promoter, SD and *cis*-regulatory sequences. By searching for motif occurrences on the complete genome of *L. plantarum* we were able to perform an *ab initio* prediction of the most commonly found intergenic elements. In addition, we observed conservation of the relative positioning and spacing of promoter, SD and the carbon responsive element (CRE) sequences (10) and used this as an added criterion for the discrimination between true- and false-positive promoter elements.

Methods:

Intergenic regions search:

The 100 nt upstream of the translation start of each gene in the genome sequence of *L. plantarum* (11) were analyzed. Sequences were orientated in the same direction as the downstream-located gene. If the 100 nt upstream sequence included coding sequences of upstream genes, the sequence was trimmed and only the non-coding sequence was used for analysis. Intergenic sequences less than 25 nt in length were not evaluated. Gene predictions were based on the 4th version of the *L. plantarum* genome annotation (July 2003).

MEME (12) was run to identify conserved motifs in the intergenic regions using DNA settings. Additional altered parameters were 1) search for Zero or One Occurrence Per Sequence (mod ZOOPS), 2) identify up to ten different motifs and 3) each motif should be found in at least thirty-five different sequences.

MAST (13) analysis was performed with the individual motifs, using either the original database of upstream sequences, or the complete chromosomal DNA sequence of *L. plantarum*. In all cases p-values were used to score significance. The level of significance was set on $1e^{-04}$.

Motif correlation

All identified motifs were used to search the database of upstream sequences. In the second step, all possible motif pairs were identified by counting the number of upstream sequences where both motifs co-occurred. These numbers of co-occurring instances were divided by the total number of hits found for the motif with the largest amount of hits to correct for a high degree of correlation between a low and high occurring motif.

Transcriptional unit prediction

Transcriptional unit (TU) prediction was based on three genome-context parameters: genes were considered to be present in the same TU if 1) adjacent genes were positioned on the same coding strand, 2) adjacent genes had an intergenic region <100 nt, and 3) no Transterm (14) predicted (Rho-independent) termination signal was present between adjacent genes.

Promoter searches

The spacer region between the -35 and -10 region of the sigA-binding motif was altered manually by either copying or deleting between the -35 and -10 regions. Rows containing the least functional information (low overall signal and low variability in the nucleotide conservation) were used for matrix alteration. MAST was performed with these new matrices to find promoter occurrences with the altered spacing region. The total amount of occurrences was, together with the hits for the extended -10 motif, filtered for overlapping promoter elements. In case of overlap, the hit with the lowest p-value was used for further analysis.

Start site comparisons

All significant elements (p-value < $1e^{-04}$) were divided in two categories: 1) inside protein-coding sequences, and 2) (partially) in the intergenic region. For all elements, the distance to the first downstream-located gene (in the same orientation) was determined. These distances were grouped and visualized in distribution plots.

Results:

In total, 2178 out of 3009 upstream regions of the genes annotated in the *L. plantarum* genome were searched for the 10 best-conserved nucleotide elements using MEME (Table 1). Residual intergenic sequences were discarded on basis of size (<25 nt; see Methods section). For all conserved sequence elements identified by MEME, a MAST search was performed on the database of upstream sequences. On basis of the MAST results, the co-localization of the different motif pairs was studied. Of the complete set of 10 motifs, the motifs 1 and 2, and 7 and 9 were found to co-occur in at least 1/3 of the upstream regions.

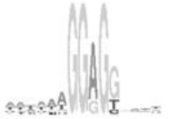
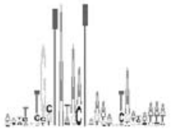
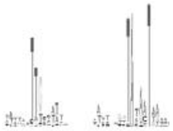
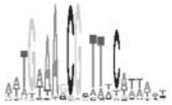
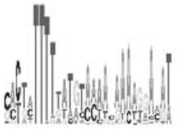
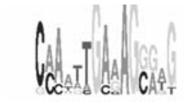
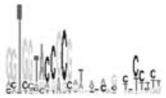
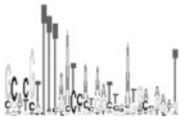
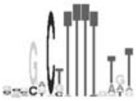
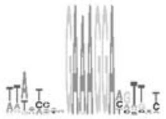
From the 10 motifs identified, at least seven resembled known elements involved in transcription or translation in different bacteria. The element with the lowest e-value (motif 1, e-value 2.7e^{-239}) displayed clear resemblance with the consensus SD sequence of bacteria (3) (AGGAGG). However, the SD sequence of *L. plantarum* (AATAAGGAGGAATT) appeared to encompass additional conserved nucleotides located directly upstream and downstream of the canonical consensus sequence (AGGAGG). These additional conserved nt form an adenine/thymidine stretch and could be related to the general C+G content of the organism (44.5 % (11)), but could also lead to a better 16S RNA recognition.

Two motifs were found that resemble the typical promoter elements of the house-keeping sigma factor, σ70 or SigA (15). Motifs 2 (e-value 1.1e^{-149}) and 3 (e-value 3.8e^{-82}) resembled the -10 box (consensus TATAAT) and the -35 as well as the -10 boxes of the canonical sigA promoter of *B. subtilis* and *E. coli* (TTGACA and TATAAT (16)), respectively. In addition, motif 3 resembled the -35 box (consensus TTGACAT), while motif 2 contained an additional conserved guanine upstream of the consensus -10 box. This alternative promoter consensus resembles the extended -10 promoter element described in *B. subtilis* (16) and *E. coli* (17). Earlier studies elucidated that one of these two alternative SigA promoters are required for effective initiation of mRNA transcription (17).

Three conserved sequence elements appeared to be part of mobile DNA elements. The most significant occurrences of motifs 6 and 10 (cut-off 1e^{-08}) were found in the upstream regions of transposases of 15 transposon elements. Using the same significance cut-off, motif 5 was found to be located upstream of transposases in 14 out of 16 cases. The two other hits identified with this sequence element were located in the upstream regions of genes encoding a valine t-RNA ligase and a hypothetical protein. In the original annotation of the *L. plantarum* genome (11), two different types of transposons were identified (ISP1 and ISP2). Further analysis of the motifs showed that motif 5 was found in the upstream region of both types of transposons, while motif 6 was unique for ISP1, and motif 10 was exclusively present in ISP2. All sequence elements identified did not overlap with the predicted transposase-recognition sites of these two transposons (18,19), but were identified in the region directly upstream of the transposase-encoding gene, and are located downstream of the transposase-recognition site.

Motifs 4, 7 and 9 resemble consensus sequences of well-known regulatory elements. Motif 4 resembles the binding site for the transcriptional regulator catabolite control protein A, CcpA, and is known as the catabolite repression element (CRE) (10). CcpA functions as a general regulatory protein for catabolite repression in most Gram-positive organisms (for a review (20)).

Table 1 Ten best scoring motifs in the upstream regions.

Number	Motif function	E-value	Seqlogo (31) representation
1	Shine-Dalgarno sequence	2.7e-239	
2	Extended -10 promoter	1.1e-149	
3	-35/-10 promoter	3.8e-082	
4	CRE binding site (CcpA)	1.3e-042	
5	Transposon-related element	1.9e-038	
6	Transposon-related element	1.1e-013	
7	Part of the T-box regulatory element	1.8e-011	
8	Unknown	1.9e-004	
9	Part of the T-box regulatory element	5.7e-003	
10	Transposon-related element	1.3e+000	

CcpA was previously predicted to regulate the expression of approximately 200 protein-encoding genes in *L. plantarum*, which was based on a motif search performed with the motif described by Siezen et al. (21). Using the motif identified in the present study we identified 276 occurrences of CRE (cutoff 1e^{-05}). This new CRE site prediction included most (~80%), but not all of the sites predicted by Siezen et al. (21). Motifs 7 and 9 resembled different parts of the T-box regulatory RNA-element, explaining their frequent co-occurrence. This box is well-described for *Firmicutes* and is known to be involved in regulation of genes encoding tRNA ligases, amino acid biosynthesis and amino acid transport (22). Analogously, of the T-box occurrences predicted here, approximately 70% were located in the upstream regions of genes belonging to these three gene classes (for details, see Chapter 5).

To the best of our knowledge, motif 8 identified here represents a novel motif that has not been described in literature before and is not present in the regulatory database of *Bacillus subtilis* (DBTBS) (23), which can be regarded as the paradigm Gram-positive organism with the best characterized regulatory network. Although motif 8 also resembles the -10 box of the SigA-binding site, one of the crucial conserved positions in the motif (the A on the second position of the -10 box) does not seem to be conserved, making it unlikely to be a alternative SigA-binding site. The annotation of the genes found to be preceded by this element (Table 2) does seem to display functional congruency. Four different genes are predicted to be involved in cell-wall related processes, of which three in peptidoglycan synthesis and degradation. In addition, genes encoding four transporters with unknown substrate, a putative β-lactamase (*lp_2385*), an aminopeptidase (*lp_0088*) and an extracellular protein with unknown function (*lp_0869*) were found to be preceded by this element. These findings support the hypothesis that motif 8 functions as a *cis*-regulatory element involved in regulation of cell-wall related processes.

The identification of the two *L. plantarum* sigA-promoter element consensuses enabled full genome searches to obtain the full complement of promoters within its genome. The two promoter-related MEME elements (motif 2 and motif 3) were used to search the complete genome of *L. plantarum*. In total, 1141 occurrences of motif 3 and 939 occurrences of motif 2 were found. Since partial overlap of these promoter-related elements was already observed (see above), double hits were curated from the list of motif 2 and motif 3 hits, resulting in a total of 1879 candidate promoter elements in the final list. More than 50 % (961) of these promoters were located in the regions upstream of protein-encoding genes in the correct orientation to drive expression of those genes, and were thus regarded as realistic candidate promoter element. The motifs used for MAST searches did not allow for variability of spacing of the (-35 and -10 boxes; see motif 3) SigA promoter elements, which is due to the fact that this variability was not included in the original position weight matrix (PWM) as provided by MEME. However, variability in the relative spacing (16-20 nt.) of these two promoter elements has been reported previously (16). Therefore, the PWM of the -35/-10 promoter sequence motif was rewritten to new matrices that allowed for spacer regions length variation from 16 to 20 nt. Searches with these more flexible promoter consensus elements resulted in an additional set of 3352 candidate promoters. Out of this set, 455 were located within the intergenic region and correctly oriented with respect to an adjacent protein-encoding gene and were added to the candidate promoter elements list, resulting in a total of 1416

Table 2: Genes preceded by the unidentified motif 8, sorted by p-value

Gene name	Product	Main class
lp_2998	integral membrane protein	Hypothetical proteins
lp_3448	unknown	Hypothetical proteins
acm1	muramidase	Cell envelope
lp_2849	ABC transporter, ATP-binding protein	Transport and binding proteins
pepI	prolyl aminopeptidase	Protein fate
lp_3433	unknown	Hypothetical proteins
lp_2599	transcription regulator	Regulatory functions
lp_2212	unknown	Hypothetical proteins
lp_0182	mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase	Cell envelope
lp_0627	prophage P1 protein 4	Other categories
lp_2452	prophage P2a protein 5	Other categories
lp_3323	unknown	Hypothetical proteins
rpiA2	ribose 5-phosphate epimerase	Energy metabolism
lp_2385	beta-lactamase (putative)	Hypothetical proteins
lp_1335	ABC transporter, ATP-binding protein	Transport and binding proteins
lp_0869	extracellular protein	Hypothetical proteins
gph1	phosphoglycolate phosphatase (putative)	Central intermediary metabolism
lp_0874	bifunctional protein: amino acid aminotransferase; 2-hydroxya-cid dehydrogenase	Energy metabolism
lrgB	effector of murein hydrolase (putative)	Cell envelope
lp_1906	unknown	Hypothetical proteins
galE1	UDP-glucose 4-epimerase	Purines, pyrimidines, nucleosides and nucleotides
lp_2625	acetyltransferase (putative)	Hypothetical proteins
lp_1686	acyl-CoA thioester hydrolase (putative)	Central intermediary metabolism
dnaX	DNA-directed DNA polymerase III, gamma/tau subunit	DNA metabolism
lp_0949	unknown	Hypothetical proteins
lp_3021	protein containing diguanylate cyclase/phosphodiesterase domain 1 (GGDEF)	Hypothetical proteins
lp_0075	acyl carrier protein phosphodiesterase (putative)	Fatty acid and phospholipid metabolism
lp_2525	ABC transporter, ATP-binding protein	Transport and binding proteins
mesJ	cell cycle protein MesJ	Cellular processes
galR1	galactose operon repressor	Regulatory functions
lp_3059	cell surface protein precursor	Cell envelope
lp_1135	prenyltransferase	Biosynthesis of cofactors, prosthetic groups, and carriers
lp_0455	transport protein	Transport and binding proteins
lp_1754	methylase (putative)	Hypothetical proteins
lp_3100	oxidoreductase	Hypothetical proteins

Sub class	**P-value**
Conserved: membrane proteins	2.60E-13
Not conserved: other	3.40E-12
Murein sacculus and peptidoglycan	8.30E-12
Unknown substrate	8.30E-12
Degradation of proteins, peptides, and glycopeptides	1.60E-11
Not conserved: other	1.90E-11
DeoR-family regulators	2.60E-10
Conserved: other	3.20E-09
Murein sacculus and peptidoglycan	8.30E-09
Phage and prophage related functions	8.30E-09
Phage and prophage related functions	8.30E-09
Not conserved: other	5.90E-08
Pentose phosphate pathway	1.30E-07
Conserved: putative function	1.30E-07
Unknown substrate	1.90E-07
Not conserved: other	2.10E-07
General	3.30E-07
Amino acids and amines	4.00E-07
Murein sacculus and peptidoglycan	4.40E-07
Conserved: other	4.40E-07
Sugar-nucleotide biosynthesis and interconversions	4.40E-07
Conserved: putative function	5.20E-07
General	1.00E-06
DNA replication recombination, and repair	1.80E-06
Conserved: other	1.80E-06
Conserved: putative function	1.80E-06
General	2.00E-06
Unknown substrate	2.50E-06
Cell division	2.70E-06
LacI-family regulators	3.20E-06
Cell surface proteins: LPxTG anchor	3.70E-06
Menaquinone and ubiquinone	4.60E-06
Unknown substrate	5.70E-06
Conserved: putative function	7.10E-06
Conserved: putative function	7.10E-06

candidate promoters. Notably, 1399 of these predicted promoters were located in the upstream regions of genes predicted to be the first gene of a transcriptional unit (TU). Overall this analysis resulted in the prediction of a candidate promoter in the non-coding upstream sequences of 72% of all predicted TUs.

To gain insight in the distribution of the conserved motifs, the distance from the end of a motif to the translation start (ATG) of the first correctly oriented downstream-located gene was analyzed (Figure 1). Motifs found at distances larger than 200 nt were not taken into account. The distribution of only three elements was analyzed (SD sequence, SigA-promoters and CRE) as the other elements were found only a few times (35-48 hits). These three elements each had more than 275 hits. As anticipated, the best correlation between gene start and motif end was observed for the SD sequence, which correlates with their position-dependent functioning in ribosomal recognition and mRNA translation. Interestingly, gene start distance correlations were also observed for the promoter and CRE elements. The *L. plantarum* promoters are preferably located within 11 to 40 nt upstream of the gene start, while CRE sites seem to share a preferred positioning somewhere between 21 and 60 nt (44% of the CRE sites were found in this region). In previous studies it was already shown that (experimentally verified) promoters have a preferred distance to translation start of the downstream gene (5). The observed spatial distribution of the *L. plantarum* promoters is in good agreement with this experimentally validated preference and thus supports the validity of the promoter prediction. Further analysis of the relation between CRE and the promoter revealed a positional relation between these two motifs, although no fixed distance between CRE and the promoter was observed in *L. plantarum*, which is in contrast with previous studies in *Bacillus subtilis* (24) and *Lactococcus lactis* (25). Nevertheless, a limited preference for CRE sites to locate at specific positions within the predicted promoter (-35 and -10 region) and the transcription start (approximately 10 nt downstream the end of the promoter, data not shown).

When comparing the frequency of intergenic versus intragenic promoters (Figure 2), an interesting observation was made. Of 967 promoters located within 100 nt of the downstream gene they are predicted to regulate, only 4.1% (40 promoter elements) are intragenic. Further analysis revealed that no intergenic promoters are found anymore between 500 and 1300 nt from the gene start. Since intergenic regions of such a size are rare within prokaryotic genomes, this is not a surprising observation. In addition, the number of distantly located promoters (>500 nt from the downstream located gene) increases continually (Figure 2), suggesting that these candidate promoter elements are most probably false-positive identifications. This observation can be used to estimate the number of false-positive promoters found within the first 500 nt. By extrapolating the continuous increase in false-positive hits, we can predict the specificity rate of the prediction. Out of 1969 predicted promoters within the first 500 nt, approximately 700 can be considered false-positive. As the number of false-positive hits should increase at a constant rate, this results in predicting 1.4 false positive hits per position (nucleotide). The ratio between the total number of hits and the number of false-positives represents the likelihood of detecting a promoter. As an example, 32 hits are found at a distance of 28 nt from the gene start. For all promoters at position 28 it can thus be concluded that the likelihood of being true-positive is 0.96 (32 – 1.4 divided by 32). On basis of this

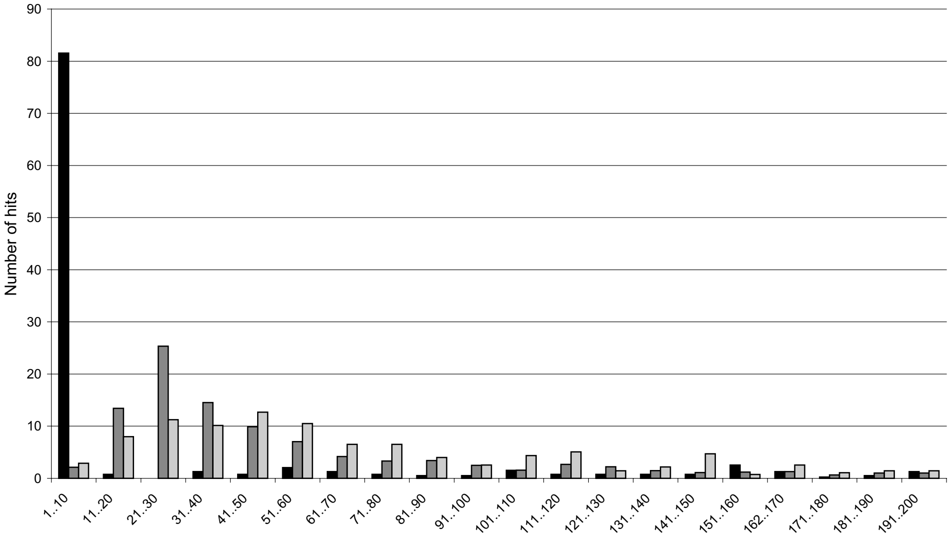


Figure 1: Distribution of four motifs in relation to the gene starts.
Black: Shine-Dalgarno sequence; dark grey: combination of the two different promoter motifs; light gray: CRE. Occurrences were grouped in subsets of ten nt.

likelihood, it can be concluded that 877 promoters (45% of the complete prediction) that end between 11 and 90 nt upstream of the translation start can be assigned as true-positive with a likelihood larger than 0.85. In addition, promoters found at distances further than 170 nt from the translation start are more likely to be false-positive than true-positive (likelihood of 0.49).

Conclusions and Discussion:

Here we show that analyzing the intergenic regions within a single genome results in identification of the most dominant and frequently used as well as the best-conserved elements. Many of these elements are involved in the primary transcription and translation processes. In the analyses presented here, the ten most prominent motifs were analyzed in more detail. However, the e-value of the worst motif (motif 10, e-value 1.3) suggests that no other highly conserved motifs will be found with altered settings. Absence of additional motifs with high abundance and conservation in the genome could correspond with the finding that only a few renowned global regulators are present in the *L. plantarum* genome. A recent study showed that *L. plantarum* indeed lacks many well-known global regulators like AbrB, CodY, ComK, Fur and TnrA (26). The binding sites for so-called local-transcriptional regulators were not identified in this study, probably because they are only represented in a few copies and could be less well conserved. To identify these elements, phylogenetic footprinting studies (7,8) are probably better suited strategies for their identification, since these methods focus on specific and preselected upstream regions. Alternatively, such *cis*-acting elements might be discovered on basis of post-genomic experimental analyses like transcriptomics.

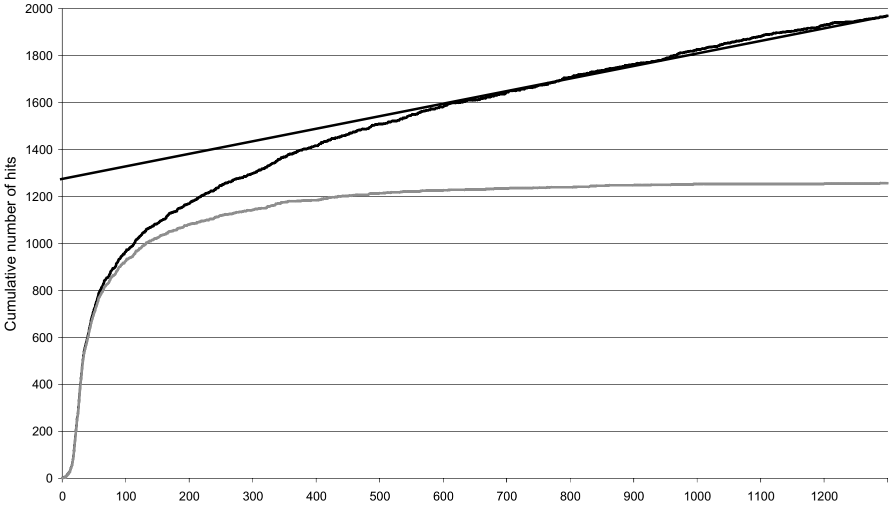


Figure 2: Cumulative number of promoters in relation to the gene starts.
Black: complete dataset; grey: intergenic promoters only. The black line indicates the slope of the total number of promoters at a distance over 500 nt to the start of the first downstream-located gene.

Three of the motifs identified here were found in the upstream regions of transposases genes. The function of these motifs remains to be established, but could, like the other motifs identified in this study, very well relate to transcription or translation. When comparing the position of these motifs with the promoter and Shine-Dalgarno sequence, we observe that motif 5 is located at the optimal position of the promoter, while the two other motifs are found at only 8 nt from the gene start, thereby overlapping the Shine-Dalgarno sequence. From literature it is known that transposases often contain degenerated promoter and Shine-Dalgarno sequences (27-29). Although these promoter and Shine-Dalgarno sequences are degenerated, they are probably conserved between different occurrences of the transposon. Interestingly, we did observe that motif 5 is conserved in both types of transposons, suggesting that there could be a relation between the origins of these two transposons.

Some of the elements were analyzed on basis of their position relative to the first downstream-located gene. The preferred location of the sigA promoter was used as a basis for accurately discriminating between true and false-positive hits. We could predict promoter sites in the upstream regions of 45% of the predicted TUs, with an estimated accuracy of 85%. Although this prediction appears to be relatively accurate, its relatively low sensitivity could still leave room for improvement, but at the same time would most likely also introduce more false-positives. Alternatively, the TUs for which no SigA promoter was predicted could be independent of the SigA-RNA polymerase complex for their transcription. It is well established that several alternative sigma factors can occur in bacteria, each with different preferred binding sites (for an overview, see (2)). The genome sequence of *L. plantarum* encodes at least two additional sigma factors; sigma 54 and sigma 30 (11), which together could regulate some of the remaining 55% of the TUs. However, since we did not find any unknown, broadly conserved motif in the upstream regions of *L. plantarum*, it seems unlikely that these sigma factors control many different TUs. Others have shown that searches with the known sigma 54 promoter of *B. subtilis* revealed only 3 hits in *L. plantarum*. Of these three binding sites only one could be experimentally validated (M. Stevens, personal communication).

An alternative explanation for the low specificity in the identification of the SigA-promoter element is the possible overlap of the promoter sequences with additional regulatory elements. In *B. subtilis* overlap of CRE with the promoter elements was also observed and it is assumed that binding of CcpA at a position overlapping the SigA-dependent RNA-polymerase target site prevents effective transcriptional initiation (30). The presence of CRE overlapping with a promoter could result in deviations from the promoter consensus. In other cases, regulatory elements can assist in RNA polymerase recruitment. It can be assumed that in both these cases the promoter is less well conserved and can remain undetected using a consensus-based search.

References:

1. **Mira, A., Ochman, H. and Moran, N.A. (2001)** Deletional bias and the evolution of bacterial genomes. *Trends Genet, 17, 589-596.*
2. **Gruber, T.M. and Gross, C.A. (2003)** Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol, 57, 441-466.*
3. **Shine, J. and Dalgarno, L. (1975)** Determinant of cistron specificity in bacterial ribosomes. *Nature, 254, 34-38.*
4. **Hannenhalli, S.S., Hayes, W.S., Hatzigeorgiou, A.G. and Fickett, J.W. (1999)** Bacterial start site prediction. *Nucleic Acids Res, 27, 3577-3582.*
5. **Huerta, A.M. and Collado-Vides, J. (2003)** Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J Mol Biol, 333, 261-278.*
6. **Robison, K., McGuire, A.M. and Church, G.M. (1998)** A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J Mol Biol, 284, 241-254.*
7. **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.*
8. **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.*
9. **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 U S A, 99, 7323-7328.*
10. **Miwa, Y., Nakata, A., Ogiwara, A., Yamamoto, M. and Fujita, Y. (2000)** Evaluation and characterization of catabolite-responsive elements (cre) of *Bacillus subtilis*. *Nucleic Acids Res, 28, 1206-1210.*
11. **Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W. et al. (2003)** Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A, 100, 1990-1995.*
12. **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.*
13. **Bailey, T.L. and Gribskov, M. (1998)** Combining evidence using p-values: application to sequence homology searches. *Bioinformatics, 14, 48-54.*
14. **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.*
15. **deHaseth, P.L., Zupancic, M.L. and Record, M.T., Jr. (1998)** RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J Bacteriol, 180, 3019-3025.*

16. **Helmann, J.D. (1995)** Compilation and analysis of *Bacillus subtilis* sigma A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucleic Acids Res, 23, 2351-2360.*
17. **Mitchell, J.E., Zheng, D., Busby, S.J. and Minchin, S.D. (2003)** Identification and analysis of 'extended -10' promoters in *Escherichia coli*. *Nucleic Acids Res, 31, 4689-4695.*
18. **Johansen, E. and Kibenich, A. (1992)** Isolation and characterization of IS1165, an insertion sequence of *Leuconostoc mesenteroides* subsp. cremoris and other lactic acid bacteria. *Plasmid, 27, 200-206.*
19. **Ito, T., Katayama, Y., Asada, K., Mori, N., Tsutsumimoto, K., Tiensasitorn, C. and Hiramatsu, K. (2001)** Structural comparison of three types of staphylococcal cassette chromosome mec integrated in the chromosome in methicillin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother, 45, 1323-1336.*
20. **Warner, J.B. and Lolkema, J.S. (2003)** CcpA-dependent carbon catabolite repression in bacteria. *Microbiol Mol Biol Rev, 67, 475-490.*
21. **Siezen, R., Boekhorst, J., Muscariello, L., Molenaar, D., Renckens, B. and Kleerebezem, M. (2006)** *Lactobacillus plantarum* gene clusters encoding putative cell-surface protein complexes for carbohydrate utilization are conserved in specific gram-positive bacteria. *BMC Genomics, 7, 126.*
22. **Grundy, F.J. and Henkin, T.M. (1994)** Conservation of a transcription antitermination mechanism in aminoacyl-tRNA synthetase and amino acid biosynthesis genes in gram-positive bacteria. *J Mol Biol, 235, 798-804.*
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, D75-77.*
24. **Kim, J.H., Yang, Y.K. and Chambliss, G.H. (2005)** Evidence that *Bacillus catabolite* control protein CcpA interacts with RNA polymerase to inhibit transcription. *Mol Microbiol, 56, 155-162.*
25. **Zomer, A.L., Buist, G., Larsen, R., Kok, J. and Kuipers, O.P. (2007)** Time-resolved determination of the CcpA regulon of *Lactococcus lactis subsp. cremoris* MG1363. *J Bacteriol, 189, 1366-1381.*
26. **Lozada-Chavez, I., Janga, S.C. and Collado-Vides, J. (2006)** Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res, 34, 3434-3445.*
27. **Simons, R.W. and Kleckner, N. (1983)** Translational control of IS10 transposition. *Cell, 34, 683-691.*
28. **Dalrymple, B. and Arber, W. (1985)** Promotion of RNA transcription on the insertion element IS30 of *E. coli* K12. *Embo J, 4, 2687-2693.*
29. **Duval-Valentin, G., Normand, C., Khemici, V., Marty, B. and Chandler, M. (2001)** Transient promoter formation: a new feedback mechanism for regulation of IS911 transposition. *Embo J, 20, 5802-5811.*
30. **Henkin, T.M. (1996)** The role of CcpA transcriptional regulator in carbon metabolism in *Bacillus subtilis*. *FEMS Microbiol Lett, 135, 9-15.*
31. **Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004)** WebLogo: a sequence logo generator. *Genome Res, 14, 1188-1190.*

CHAPTER 4

Large Intergenic Cruciform-Supermotifs in the Lactobacillus plantarum Genome: a putative RNAi-like regulatory role?

Michiel Wels
Roger S. Bongers
Jos Boekhorst
Douwe Molenaar
Mark Sturme
Willem M. de Vos
Roland J. Siezen
Michiel Kleerebezem

Submitted for publication

Acknowledgements

The authors are grateful to Mariela Serrano, Arno Wegkamp, Marc Stevens and Rob Brooijmans for providing their microarray data prior to publication for transcriptome meta-analysis, to Christof Francke for insightful discussions and John van der Oost for critically reading the manuscript.

All supplementary files regarding this chapter can be found at www.cmbi.ru.nl/~mwels/Chapter_4

Conserved DNA sequence motifs were searched in large (>700 nt) intergenic regions of the *L. plantarum* WCFS1 genome sequence. Several highly conserved, small motifs (8 – 34 nt) were found that are genetically linked and form multiple mosaic *Lactobacillus plantarum* supermotifs (LPSM). A Hidden Markov Model (HMM) was used to refine LPSM detection, revealing 24 LPSM occurrences with a length varying from ~800 to 1000 nt. Secondary structure analysis predicted conserved cruciform-like structures for these LPSM. The LPSMs appear to be unique for *L. plantarum* and were found to be highly conserved among different *L. plantarum* strains by DNA microarray analyses and locus-specific PCR amplification and sequencing. Northern blot analysis and meta-analysis of transcriptome datasets indicated transcription of these LPSMs. In addition, transcriptome data suggested regulation of the expression in experiments comparing *L. plantarum* WCFS1 wild type with a strain overexpressing the endogenous thioredoxin reductase encoding *trxB1* from an expression plasmid. Intriguingly, similar regulatory effects were observed in other studies comparing *L. plantarum* WCFS1 and strains overexpressing genes *in trans*. Although the function of these LPSMs remains unknown, they could play a role as regulatory RNAs acting as a defense system against nucleotide presence or expression that could be detrimental to the cell.

Introduction

Intergenic regions of bacterial genomes often are considered to play a key role in the expression regulation of the flanking coding regions and contain *cis*-acting sequences that act as promoters, transcription or translation regulation binding sequences or terminators. These regulatory elements are generally located in relatively short intergenic sequences (mean size of 147 nt compared to 947 nt for coding genes) (1). However, occasionally much larger non-coding regions are encountered in bacterial chromosomes. In some instances these large non-coding regions appear to contain additional functional genetic elements. Most of these elements are part of the 5'-untranslated region (UTR) of the mRNA and are involved in premature transcription termination. Examples of this are riboswitches and T-box elements (2,3). In addition, other larger non-coding elements appear to function as *trans*-acting RNA elements. The encoded RNA molecules of the latter group are not translated into protein, but can bind a regulatory target, thereby blocking its transcription or translation. Although the majority of such *trans*-acting regulatory RNAs is relatively short (generally up to 100 nt) (4), some examples of much larger RNAs have been described. A well-studied example is the 514 nt RNA III in *Staphylococcus aureus* (5), which has a key role in the complex regulation circuits that control virulence gene expression in this species (for a review, see (6)).

In addition to these established *trans*-acting regulatory RNA encoding sequences, there are examples of repeated elements in bacterial chromosome that have an undefined function. In *Bacillus subtilis* a 190 nt repeated sequence was identified in 10 different intergenic regions (7), and a comparable element was found in *Mycobacterium leprae* (8). Although the function of these motifs remains unknown to date, they have been suggested to originate from and represent remnants of IS-like elements. The largest family of repeated elements described to date has been identified exclusively in *Staphylococci* and has been designated *Staphylococcus aureus* repeat (STAR) element (9). STAR elements consist of different numbers of internally repeated DNA sequences and have been identified in two conserved loci in twelve different strains. In general, STAR elements have a length of 300 – 600 nt, but occasionally larger STAR elements are found. The largest STAR element characterized to date has a size of 1,127 nt. Southern blot analysis showed the possible presence of tens of STAR-element copies in each *Staphylococcus* strain (9). Sequencing analysis showed that the number of repeats differs per strain, and can therefore be exploited for strain-specific genotyping (10). Another smaller region (BOX-element, 100-200 nt) consisting of internally repeated sequences has been identified in *Streptococci.* 127 BOX-elements were found in the *Streptococcus pneumoniae* TIGR4 genome, but were absent in *Lactococcus lactis* IL1403 and *Streptococcus pyogenes* M1 GAS(11). BOX-elements have also been proposed to provide an amplification target that would be suitable for strain-specific fingerprinting genotyping efforts (12).

Here we describe the identification of a novel sequence repeat (designated *L. plantarum* supermotif (LPSM)) present in multiple copies in the intergenic regions of the genome of *L. plantarum* WCFS1. The identified LPSM element contains many inverted repeat regions and is predicted to fold into cruciform-like, low-energy secondary structures. While these motifs appeared to be unique for *L. plantarum*, they were consistently found in different strains of this species. Intriguingly, Northern blot analysis and DNA microarray-based transcriptome profiling support that the LPSMs are transcribed into RNA. Nevertheless, a clear biological function of these repeats or their transcripts remains to be determined.

Material and Methods:

Bacterial strains and culturing conditions

All strains, plasmid and primers used in this study are listed in tables 1 and 2, respectively. *L. plantarum* was grown at 37°C in MRS broth (Difco, Surry, UK) without aeration. Stationary phase incubation was performed for fourteen days at 20°C in MRS broth without aeration. *E. coli* was grown in Tryptone Yeast (TY) medium (33) with aeration at 37°C. Where appropriate, ampicillin was added to the medium at a concentration of 100 µg ml^{-1}.

DNA isolation, manipulations and sequence analysis

JetStar columns (Genomed, Oberhausen, Germany) were used for large-scale isolations of *E. coli* plasmid DNA following the instructions of the manufacturer. Small-scale plasmid DNA isolations and standard recombinant DNA techniques were performed as was previously described (33). Plasmid transformation of *E. coli* cells was perfumed using the $CaCl_2$ procedure (33). Chromosomal DNA of *L. plantarum* was isolated as described previously (34). DNA was isolated from agarose gels by using a QIAEX II gel extraction kit (Amersham Pharmacia Biotech, Roosendaal, The Netherlands). Primers annealing to sequences within the genes flanking eight different LPSMs were designed (Table 2) and used to amplify the intergenic regions in 3 other *L. plantarum* strains (Lp_80, ATCC 14187, 299v). PCR amplifications were carried out with an automated thermal cycler (Perkin-Elmer, Shelton, Conn.) using *Taq* DNA polymerase (Gibco-BRL Life Technologies, Breda, The Netherlands). The PCR amplicons were cloned in pGEM-T using the pGEM-T Easy Vector Systems (Promega Corp., Madison, WI). pGEM-T insert sequencing was performed by BaseClear (Leiden, The Netherlands) using standard pUC-based forward and reverse sequencing primers. The sequences obtained were aligned

Table 1: used strains and plasmid

Strain (source/reference)	Relevant characteristics / reference / source
E. coli DH5α	(44)
L. plantarum WCFS1	Sequenced strain (14)
L. plantarum 299V	(45)
L. plantarum lp_80	(17)
L. plantarum ATCC14917 (ATCC)	ATCC
Plasmid (source)	
pGEM-T	PCR cloning vector; ApR (Promega)

Table 2: Primers used

Primer	Sequence
M2 forward	CCGGAGCAGCAACCGCAACCATGATG
M2 reverse	TCTCAGTCCCTTCGCTAGAATCGC
M3 forward	GTTGGCAACGTGGCTGATGGGG
M3 reverse	GGGACAGTGCTATGATTGTCGCACTCC
M4 forward	CGACGCGACGCTGAATGCCGATCC
M4 reverse	CCGCCCTGGAAGAACAGATGACTGCC
M11 forward	CCACGGTCGCAATTTGTAAGGACTG
M11 reverse	CAATGGGAATTCCTGTTATCGTTGGTG
M11 internal 1	CGTCGATTTGAGCTCACGCAG
M11 internal 2	GGTGCCAAAACAATTGAGAC
M14 forward	CATGTTGAGTGGTCCAAGCACCCG
M14 reverse	GCCAAGTTACCGACCTGCGAG
M16 forward	CGAGAACAGATGTTGCACCGAC
M16 reverse	GCGGCGCCTGAGGGTGTTAAGGC
M26 forward	CCTTGATCTCATCAAGCACGCGACGC
M26 reverse	GCGGCCACCGCTTGCCCACAG
M26 internal 1	GAGCTAACGCCCAACCCGCG
M26 internal 2	CAGAGGGCTTAATACACCTAA
M32 forward	CCGCCCAACGCCCCTGAAAG
M32 reverse	CCCAAGGACAATCAGAATCCCCG
M32 internal 1	TCGATTTGAGCTAACGCACAA
M32 internal 2	GCAACTGTCAGGTAGAAGGTA

with the corresponding *L. plantarum* WCFS1 sequences using Muscle (35).

RNA isolation and Northern blotting

L. plantarum grown in MRS medium to an optical density at 600 nm (OD_{600}) of 3. Cells were harvested by centrifugation and total RNA was isolated using the macaloid method (36) including the adaptations described previously (37). Northern blotting was performed as was previously described (38). The LPSM 2 specific amplicon (primers M2 forward and M2 reverse; see Table 2) obtained using *L. plantarum* WCFS1 chromosomal

DNA as a template for amplification was used as a probe and was labeled with [α-^{32}P]dATP using nick-translation (39).

Bioinformatics:

MEME

The intergenic regions on the chromosome of *L. plantarum* WCFS1 were identified using the ORF predictions as found in Genbank. All intergenic sequences larger than 700 nt were analyzed. MEME (13), an Expectation Maximization based algorithm, was used for motif detection, with the following settings: model: TCM; minimum number of occurrences: 20; width range between 10 –100 nt and a maximum of 15 different motifs. The best hits were manually inspected for genomic co-occurrence in the original set of intergenic regions. To get a full list of the co-occurring motifs in *L. plantarum*, a MAST (40) was performed on the total chromosomal sequence. MAST was run using default program settings.

Hidden Markov Model (HMM) searching

All stretches of 10 or more co-occurrences of the different MEME motifs were aligned using Muscle (35) and built into a LPSM-specific HMM. A search with this HMM was performed on the complete chromosome of *L. plantarum* WCFS1 using HMMer (41). All HMM hits with an e-value < 0.01 were considered statistically relevant. These hits were built into a new HMM and used to search iteratively against the chromosome of *L. plantarum* until no additional hits were identified. To test the specificity of the model, an HMMer search was performed on the complete inverted chromosomal sequence (3' – 5'), which did not result in hits with an e-value < 0.1. In addition, all hits were randomly scrambled and compared against the HMM. None of the scrambled sequences fitted the HMM with an e-value <1. The HMM was used to search all publicly available completely sequenced genomes gathered from the NCBI repository (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) on November 1, 2006. In addition to the HMM searches, the genomes were also scanned using MAST-based searches with the original MEME motifs.

Secondary structure prediction

The Mfold program (42) was used for secondary structure prediction. All structures were manually compared and divided into different sub-families on basis of their predicted structure.

Design and meta analysis of microarrays

Sixty-mer oligonucleotides were designed and spotted on Agilent microarray chips (43). Only one of the LPSMs was spotted on the microarray with complete coverage. Consecutive oligonucleotides were designed in a 20 nt overlap tiling manner to cover the complete LPSM2 sequence resulting in 16 probes per DNA strand. In addition to the full coverage of LPSM2, all other LPSMs were represented by two oligonucleotides corresponding to the two best conserved 60 nt sequences within the family of LPSMs. For both approaches, both plus- and minus-strand sequences were spotted on the array. Using all 72 different *L. plantarum* transcriptome datasets available in our microarray database on March 2nd 2006, a transcription meta-analysis was performed. Mean expression levels were calculated per experiment for each oligonucleotide corresponding to part of the LPSM. Expression levels were corrected for signal background by local background signal subtraction.

Results:

Scanning intergenic DNA regions

To identify conserved genetic elements in the intergenic regions of *L.* plantarum, all 74 non-coding regions larger than 700 nt were scanned for conserved sequence motifs using MEME (13). This analysis revealed twenty different motifs with a length varying from 19 to 100 nt. The largest of these repeated sequences (100 nt) was found to correspond to eight small open-reading frames that were missed during the initial annotation process of the *L. plantarum* genome (14). Remarkably, the other 19 motifs (all smaller than 50 nt) were found to co-occur in the same intergenic loci, indicating the existence of a large conserved intergenic sequence in *L. plantarum*. Indeed, careful inspection of the conserved order of these motifs showed these to form a > 800 nt supermotif, which is termed here *L. plantarum* supermotif (LPSM).

Chromosome-wide detection of the LPSM

To identify the sequence structure of the LPSM, co-occurrence of the smaller motifs was manually scrutinized for all 74 initially used intergenic regions. Twenty-six of these 74 regions were found to have at least ten consecutive motifs in one stretch of the chromosomal sequence (with less than 50 nt in between each consecutive motif). A Hidden Markov Model (HMM) was built from the complete alignment of these 26 regions and used to search the complete genomic sequence of *L. plantarum* WCFS1. Hits below threshold (E-value of 0.1) were used to build a refined HMM. This procedure was iterated until no additional hits were found. In total, 66 LPSM hits were identified with an E-value below threshold. All minus-strand hits exactly co-localized with a plus-strand hit, indicating that there are 33 copies of a conserved bi-directional LPSM within the chromosome of *L. plantarum* WCFS1 (for an alignment of the 33 copies see supplementary Fig. S1). Motifs were numbered on the basis of their position on the chromosome. The 7 copies found in addition to the 26 originally identified, partially overlapped with potential open reading frames (ORFs), explaining their lack of detection during the MEME-based search that was restricted to intergenic regions. These LPSM-overlapping ORFs may encode small proteins (< 100 AA) that have no significant homologues in any other organism. Hence, we assume that all LPSMs are located in intergenic regions. On basis of the alignment a maximum likelihood phylogenetic tree was constructed using PHYML (15) (supplementary Fig. S2). For each of the 33 motifs, the branch length to the last branch point was taken as a measure for uniqueness. A complete list of LPSM positions, the best score with the HMM (including the direction of the best hit) and their sequence conservation to other LPSM sequences can be found in the supplementary material (Table S1). The highest similarity to the HMM was found for LPSM 26 (0.026), while the least conserved motif appeared to be LPSM 9 (0.612). The worst-scoring LPSM sequences are not all part of one deviating subfamily, but are randomly distributed (probably false-positive) hits. Nine LPSM sequences, with a uniqueness score over 0.150, were regarded as false-positive. These false-positive hits also had the worst scores against the HMM (Table S1).

To obtain additional insight in a possible role of these LPSMs, they were analyzed in relation to their chromosomal context and their relative positioning in correspondence with chromosome-wide sequence-characteristics. The 33 LPSM sequences appeared to be randomly distributed along the *L. plantarum* chromosome (Fig. 1). Moreover, the G+C content of the LPSMs (41 to 47%) was found to deviate not significantly from the average of the complete genome (44.5%) (14). In addition, the G+C content, GC-skew and codon adaptation index (CAI) of directly flanking genes of the LPSMs do not differ from other parts of the chromosome (Fig. 1).

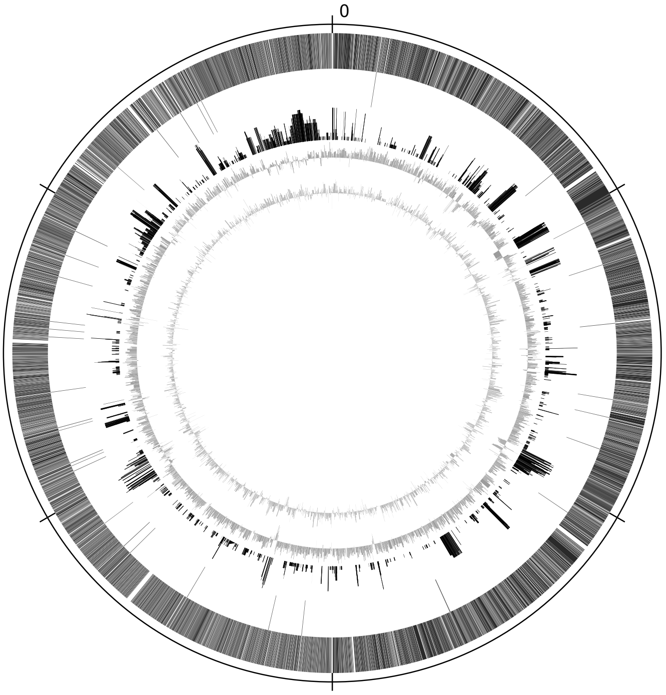


Figure 1 (a color version of the figure can be found at page 168): Chromosome wheel displaying the identified LPSM hits on the chromosome of *L. plantarum*.
Outer ring (ring 1): predicted ORFs (+ and – strand, blue and red, respectively), ring 2: LPSM occurrences, ring 3: Relative conservation among 20 *L. plantarum* strains as determined by array-based chromosome profiling. Peak height represents the number of strains, in which specific chromosomal regions of strain WCFS1 were scored as absent in other strains of L. plantarum (17), ring 4: G+C content centered around the median G+C content, ring 5: Codon Adaptation index. Figure generated using MGV (46).

For all LSPMs, the genomic context and flanking gene order was compared with other related bacteria. Seventeen out of 33 LPSMs were found in regions that share a similar gene order (some examples are displayed in Fig. 2) with other genomes, in particular other lactic acid bacteria (16). The corresponding intergenic regions in *L. plantarum* were generally longer, indicating that an LPSM is uniquely present in these loci in the *L. plantarum* genome.

To detect whether the identified LPSM is present in other bacteria, the HMM was used to scan all publicly available genomes including the recently sequenced LAB (16). These results indicate that the full length LPSM is uniquely found in *L. plantarum*. To determine whether the LPSMs are conserved among *L. plantarum* strains, the loci were analyzed using the data from a previously published array-based *L. plantarum* genomic diversity study (17). Based on these comparative genome hybridization (CGH) data there was no reason to assume that the different LPSMs were absent in any of the tested 20 *L. plantarum* strains (Fig. 1, ring 3). However, since the LPSM sequences in *L. plantarum* WCFS1 show a high level of sequence conservation, cross-hybridization cannot be excluded. In addition the CGH analysis does not disclose information with regard to the chromosomal positioning of the LPSMs in other strains. Hence, a PCR-approach was

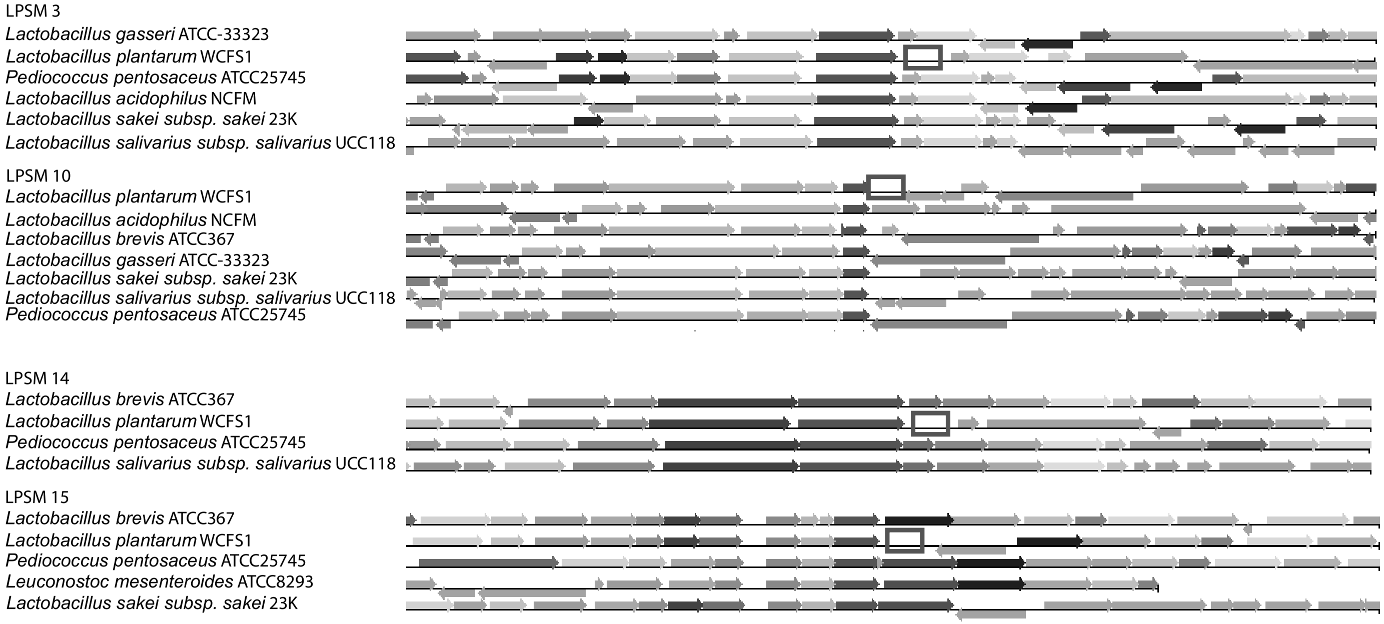

LPSM 3
Lactobacillus gasseri ATCC-33323
Lactobacillus plantarum WCFS1
Pediococcus pentosaceus ATCC25745
Lactobacillus acidophilus NCFM
Lactobacillus sakei subsp. sakei 23K
Lactobacillus salivarius subsp. salivarius UCC118
LPSM 10
Lactobacillus plantarum WCFS1
Lactobacillus acidophilus NCFM
Lactobacillus brevis ATCC367
Lactobacillus gasseri ATCC-33323
Lactobacillus sakei subsp. sakei 23K
Lactobacillus salivarius subsp. salivarius UCC118
Pediococcus pentosaceus ATCC25745
LPSM 14
Lactobacillus brevis ATCC367
Lactobacillus plantarum WCFS1
Pediococcus pentosaceus ATCC25745
Lactobacillus salivarius subsp. salivarius UCC118
LPSM 15
Lactobacillus brevis ATCC367
Lactobacillus plantarum WCFS1
Pediococcus pentosaceus ATCC25745
Leuconostoc mesenteroides ATCC8293
Lactobacillus sakei subsp. sakei 23K

Figure 2 (a color version of the figure can be found at page 169): Genome context of six LPSM locations in *L. plantarum* WCFS1 compared to other lactic acid bacteria. LPSM position in *L. plantarum* is denoted by the red box. Orthologous genes in different organism have the same color. Figure generated using the ERGO Bioinformatics suite (47).

used to analyze possible LPSM conservation among *L. plantarum* strains in more detail. Primers were designed to specifically anneal to genes that flank the WCFS1 LPSMs and successfully used to amplify eight different LPSM regions in the WCFS1 genome. Subsequently, these primers were used in an attempt to amplify the corresponding regions of in the chromosomal DNA isolated from three other *L. plantarum* strains (Table 1). A PCR product of a comparable size relative to the corresponding WCFS1 locus was obtained in all but one case (Fig. 3), supporting the strain and context conservation of the LPSM as suggested by the CGH study. Eight randomly picked LPSM amplicons obtained from these strains were sequenced and displayed high level (> 88%) sequence identity with the corresponding region found in strain WCFS1. In addition, a maximum likelihood phylogenetic tree of the sequenced LPSMs in combination with the 33 identified *L. plantarum* WCFS1 LPSMs showed clustering of seven out of eight of the sequenced LPSMs with their counterparts in *L. plantarum* WCFS1 (supplementary Fig. S2), suggesting that the motifs were present in the last common ancestor of the tested *L. plantarum* strains. In one case (motif 32 in *L. plantarum* 299v), internal reorganization relative to WCFS1 was observed within the 299v-specific LPSM structure (inversion of an internal part of the LPSM), suggesting structural instability of the LPSMs. This observation could be a reflection of a putative functional role of the LPSMs that may act as hotspots for chromosomal recombination, which would be in agreement with their high degree of sequence similarity. Such motif-based chromosomal rearrangements could play a role in genome plasticity.

Secondary structure prediction of the LPSM

Triggered by the finding of a high degree of internal and inverted repeat elements within the LPSM regions, the secondary structures of these LPSM DNA regions were predicted. A high level of secondary structure was predicted for all 24 LPSM sequences, with free energy values (ΔG) ranging from -93 kJmol^{-1} to -149 kJmol^{-1} and an average value of –132 kJmol^{-1}. The significance of the predicted secondary structures was confirmed by comparison with scrambled LPSM sequences, generating a mean ΔG value of -73 kJmol^{-1}. Mfold predicted similar secondary structures for 24 of the LPSM sequences, which could not be predicted from the scrambled sequences. Eighteen LPSMs displayed typical secondary structures (Fig. 4A). Notably, these structures share remarkable similarity with DNA structures with unknown function of eukaryotic origin called “cruciforms” (18). Next to the typical cruciform structure displayed by most LPSMs, some structural variations were predicted for some motifs. Two LPSM sequences (LPSM 5 and 15) fitted less well with the cruciform structure and appeared to contain an additional hairpin loop in the top of the cruciform structure (Fig. 4), resulting from a sequence insertion of

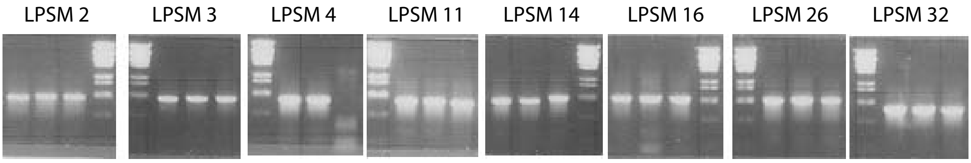


Figure 3: Agarose gel analysis of the PCR products of the *L. plantarum* strains ATCC14917, 299V and LP80, respectively. PCR reactions were performed on eight different LPSMs (numbers indicated in the figure) using purified total genomic DNA as a template.

approximately 80 nt in the centre of the LPSM sequence (supplementary Fig. S1). In addition, three motifs (LPSM 14, 20 and 33) were characterized by the lack of "arms", leading to a single-hairpin structure (Fig. 4C), and corresponding with large gaps in the multiple sequence alignment in the regions 550 to 600 and 950 to 1020 (supplementary Fig. S1). Finally, only one LPSM sequence (LPSM 28) did not appear to fit with the proposed conserved cruciform secondary structure. This LPSM sequence displayed a high ΔG value (-156 kJmol^{-1}) suggesting that this LPSMs folds into a functional structure. The structures of the false positive hits (as denoted in supplementary table S1) are not folding into cruciform-like structures. To validate the structure predictions generated by Mfold, an alternative structure prediction software tool was employed (FreeBee; (19)). This package was designed to detect conserved RNA structural features and employs a sequence alignment file as input rather than a single sequence. Nevertheless, the generic LPSM structure predicted by FreeBee analysis of the multiple alignment file (supplementary Fig. S1) strongly resembles the structure predicted for individual LPSM sequences using Mfold (data not shown). Taken together, the LPSM sequences identified in the *L. plantarum* WCFS1 genome sequence appear to have a common intrinsic folding tendency, which in most cases leads to a predicted cruciform secondary structure with a free energy value below -100 kJmol^{-1}.

The predicted common fold of the LPSM sequence elements provided an alternative structure-based rather than sequence-based method for searching similar LPSM structures in other bacterial genomes. Therefore, the intergenic regions over 500 nt of all completely sequenced genomes (see methods, HMM searching for source) were subjected to Mfold structure prediction and the results were scanned for ΔG values comparable to the values measured for the LPSM structures in *L. plantarum*. No comparable ΔG values were observed in any of the intergenic regions of the selected species, which corroborates the lack of LPSM-like sequences in other species.

Transcription of the LPSM

It is unlikely that the LPSM sequences encode proteins since they contain a many stop codons in all six reading frames on both positive and negative strands of the DNA. Nevertheless, these sequences could be transcribed as RNA molecules that potentially play a regulatory role. Northern blot analysis was performed to evaluate whether the LPSM is transcribed into RNA. A probe specific for LPSM2 detected a range of RNA products of various sizes: 2.8 kb, 1.7 kb, 1.5 kb and 0.7 kb. While the smallest transcript detected (0.7 kb) could correspond to the LPSM2 alone (788 nt), the larger transcripts most likely derive from transcripts of the LPSM2 flanking protein-coding genes that extend into the LPSM2 region. For example, a transcript of LPSM2 together with two downstream located genes (*acpS* and *alr*) would generate a product of approximately 2.8 kB, which corresponds to the size of one of the transcripts detected using the LPSM2 probe.

To further substantiate transcription of the LPSM sequence, and to compare transcription under a variety of conditions, a meta-analysis was performed on transcriptome data of *L. plantarum* WCFS1 obtained with oligonucleotide-based DNA micro-arrays that contain probes corresponding to LPSM sequences (based on LPSM 2). Although a high degree of variation in mean expression level between experiments was observed for the different oligonucleotides, it appeared that all 32 oligonucleotides gave a signal that clearly exceeds background levels in all 72 experiments. The highest expression

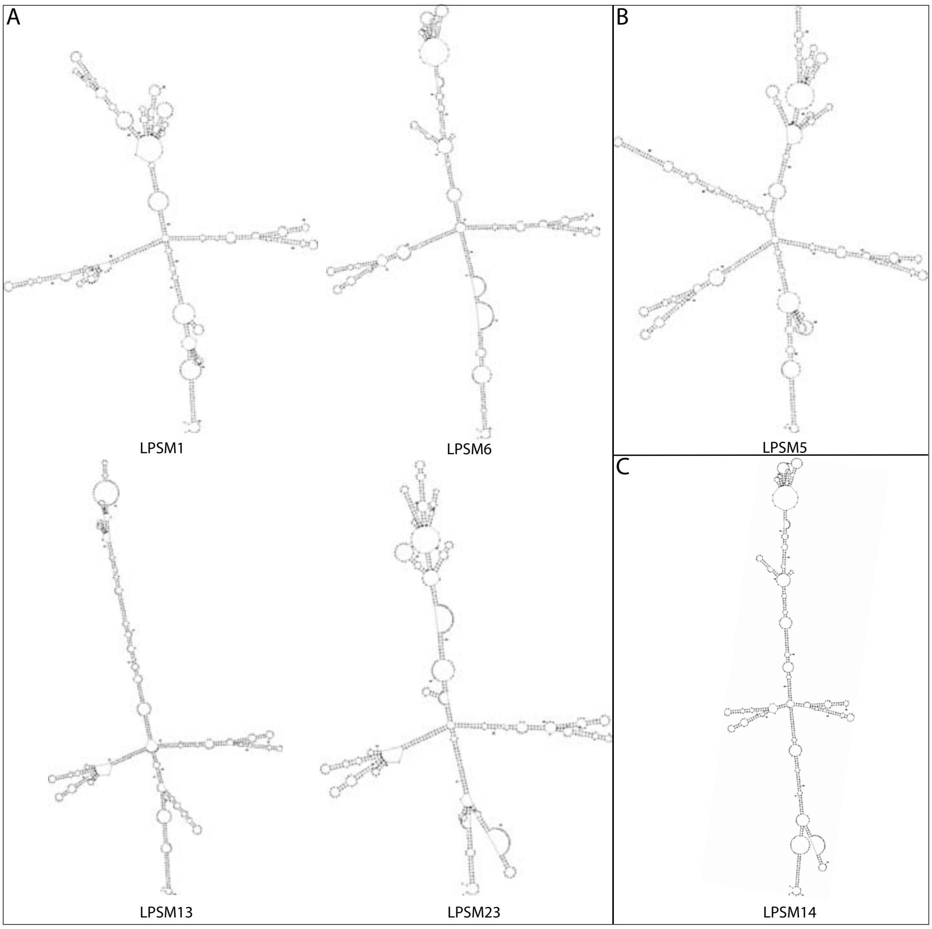


Figure 4: Exemplary two-dimensional structure predictions for LPSM sequences.
The most commonly found motif resembles the structure of a cruciform (A). Other structures (B – C) clearly resemble the cruciform structure, but have certain specific variations; structure B contains an additional hairpin in the “head” region of the LPSM structure, while C lacks the “arms” of the LPSM structure.

levels per probe were found in an experiment describing a comparison of *L. plantarum* WCFS1 versus a strain harboring one of the endogenous thioredoxin reductases (*trxB1*, *lp_0761*) on an overexpression vector as described by Serrano et al. (20) (Fig. 5). In this particular experiment, both wild type as well as the overexpression strain were subjected to H_2O_2 stress. Further analysis of the measured signals under these specific conditions showed that the high expression levels were related to a change in ratio between the two conditions (Fig 5). Oligonucleotides for which a high expression signal was measured (oligo 5; A value of 13.56) also show a high ratio between the two tested conditions (oligo 5; M value of 1.63). In all cases the expression ratios indicated increased expression level of the LPSM in the strain overexpressing *trxB1*. Furthermore, other transcriptome experiments using the same *trxB1* overexpression mutant under different conditions (without

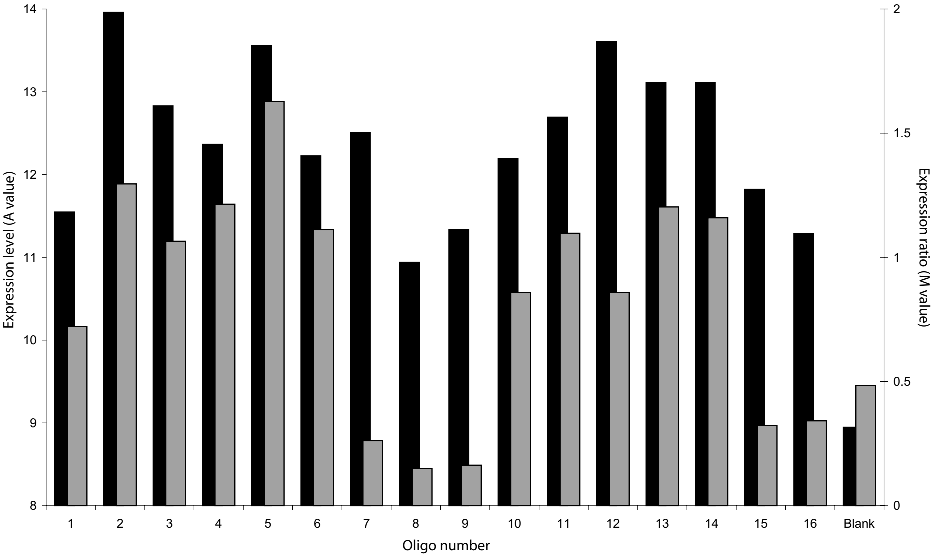


Figure 5: Expression levels (A: (log2Cy3 + log2Cy5) / 2) (black bars) and expression ratio (M: log2Cy3 - log2Cy5) (grey bars) for the oligonucleotides on the + strand of LPSM 2 under the conditions described by (20).
For the probes on the – strand, comparable signals were found.

applying H_2O_2 stress) confirmed that *trxB* overexpression correlates with induction of LPSM induction. Intriguingly, high differential levels of LPSM expression were consistently seen in various experiments comparing *L. plantarum* wild type and overexpression variants, suggesting a more general role for the LPSM transcripts in response to *in trans* overexpression of (endogenous) genes, which possibly plays a role in the defense against the effects of potentially detrimental products, as was reported by Serrano et al. (20). These meta-analysis results confirm the presence of a transcript corresponding to the LPSM sequence, and suggest that differential transcript levels are present for different regions of the LPSM. In addition, the level of the LPSM derived transcripts appears to be modulated by different experimental conditions.

Discussion

Here we describe the identification of large and unique intergenic LPSMs that are present within the chromosome of *L. plantarum* WCFS1 and other strains of the same species. The conserved LPSM is predicted to display a high intrinsic tendency to form cruciform secondary structures with high free energy levels. This could be especially relevant in combination with the observation that (regions of) these LPSM sequences appear to be transcribed differentially, as was disclosed by meta-analysis of transcription profiling datasets. From all non-coding RNA molecules described in recent literature, most are proposed or have been shown to fulfill a role in transcription regulation (21). Therefore, it is tempting to speculate that the LPSM-derived RNA molecules are involved in regulation. However, apart from the earlier described RNA III of *Staphylococcus aureus* (5), regulatory RNA molecules of the size described here and with such significant structural features are rarely described for prokaryotes. Nevertheless, our results indicate that at least part of the LPSM sequence is transcribed into RNA. It could also be that an initial transcript encompassing the entire LPSM sequence is post-transcriptionally processed into smaller RNA(s) that functions as a regulatory RNA molecule.

In eukaryotes, structures of a comparable size and with similar cruciform-like structure have been found by an *in silico* RNAi detection study (22). In these molecules, the functional interfering RNA is embedded in a large DNA structure. Part of the DNA structure functions as a regulatory element, while the lower stem is transcribed into double stranded RNA. The process of RNA interference was first described in *C. elegans* and *D. melanogaster* (23,24) and subsequently shown to function in many other eukaryotic organisms, including higher vertebrates such as mice (25). Recently, the existence of an RNAi-like process was proposed in prokaryotes, related to defense against phage or plasmid integration (26). These RNAi-like processes are linked to the existence of the Clustered Regularly Interspaced Short Palindrome Repeats (CRISPR) regions (27). RNA transcripts and post-transcriptional processing of the CRISPR transcripts have been shown in *Archaeoglobus fulgidis* (28) and *Sulfolobus solfataricus* (29). It is suggested that CRISPRs are activated by a group of proteins known as CRISPRs associated (*cas*) genes (30). *Cas* genes have been shown to occur in the direct vicinity of the CRISPR regions and employ functions related to the activity and processing of the CRISPRs (26,30). CRISPRs and *cas* genes are found in all sequenced archeal species and bacteria of all lineages, but appear to be absent in *L. plantarum* (31). Recently it was shown that CRISPR play a role in the defense against phage infection in the lactic acid bacterium species *Streptococcus thermophilus* (32). The highest differences in expression of the LPSM2 in *L. plantarum* as revealed by micro-array meta-analysis appeared to correlate with *in trans* overexpression of (in this case, endogenous) genes in *L. plantarum,* suggesting that LPSM expression is part of the response against potentially detrimental genes or gene products. Analysis of the gene neighborhood of the different LPSMs revealed the presence of several different DNA/RNA processing and repair genes located adjacent to LPSM 2, 3, 7, 12 and 14. Although some of them (e.g. *rpoD* and *rpoC*) do not seem to play a specific role in LPSM processing or activity, other genes, like *rhe1* and *radA* could be involved in processing of LPSMs, even more since the genome of *L. plantarum* encodes multiple copies (paralogs) of these putative RNA/DNA repair genes. In addition, purine and pyrimidine biosynthesis and transport genes (*upp, purP pyrR1* and *pucR*) are encountered adjacent to LPSM 16 and 22. These genes could be related to LPSM functioning in response to a change in nucleotide levels in the cell.

On basis of the presence of RNA processing enzymes in the vicinity of the LPSMs in combination with the observed transcription we hypothesize that the identified LPSMs function as a regulatory RNA molecule. As transcription of LPSM2 observed upon the *in trans* overexpression of (endogenous) genes we hypothesize that LPSM transcripts act as regulatory RNAs, involved in defense against high copy presence of certain genetic elements, especially when their expression could have detrimental effects on cell physiology in *L. plantarum*.

References:

1. **Mira, A., Ochman, H. and Moran, N.A. (2001)** Deletional bias and the evolution of bacterial genomes. *Trends Genet, 17, 589-596.*
2. **Winkler, W.C. (2005)** Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr Opin Chem Biol, 9, 594-602.*
3. **Grundy, F.J. and Henkin, T.M. (2003)** The T box and S box transcription termination control systems. *Front Biosci, 8, d20-31.*
4. **Chen, S., Lesnik, E.A., Hall, T.A., Sampath, R., Griffey, R.H., Ecker, D.J. and Blyn, L.B. (2002)** A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems, 65, 157-177.*
5. **Novick, R.P., Ross, H.F., Projan, S.J., Kornblum, J., Kreiswirth, B. and Moghazeh, S. (1993)** Synthesis of *staphylococcal* virulence factors is controlled by a regulatory RNA molecule. *Embo J, 12, 3967-3975.*
6. **Novick, R.P. (2003)** Autoinduction and signal transduction in the regulation of *staphylococcal* virulence. *Mol Microbiol, 48, 1429-1449.*
7. **Moszer, I. (1998)** The complete genome of *Bacillus subtilis*: from sequence annotation to data management and analysis. *FEBS Lett, 430, 28-36.*
8. **Woods, S.A. and Cole, S.T. (1990)** A family of dispersed repeats in *Mycobacterium leprae*. *Mol Microbiol, 4, 1745-1751.*
9. **Cramton, S.E., Schnell, N.F., Gotz, F. and Bruckner, R. (2000)** Identification of a new repetitive element in *Staphylococcus aureus*. *Infect Immun, 68, 2344-2348.*
10. **Quelle, L.S., Corso, A., Galas, M. and Sordelli, D.O. (2003)** STAR gene restriction profile analysis in epidemiological typing of methicillin-resistant *Staphylococcus aureus*: description of the new method and comparison with other polymerase chain reaction (PCR)-based methods. *Diagn Microbiol Infect Dis, 47, 455-464.*
11. **Mrazek, J., Gaynon, L.H. and Karlin, S. (2002)** Frequent oligonucleotide motifs in genomes of three *streptococci*. *Nucleic Acids Res, 30, 4216-4221.*
12. **Martin, B., Humbert, O., Camara, M., Guenzi, E., Walker, J., Mitchell, T., Andrew, P., Prudhomme, M., Alloing, G., Hakenbeck, R. et al. (1992)** A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. *Nucleic Acids Res, 20, 3479-3483.*
13. **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.*
14. **Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W. et al. (2003)** Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A, 100, 1990-1995.*
15. **Guindon, S. and Gascuel, O. (2003)** A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol, 52, 696-704.*
16. **Makarova, K., Slesarev, A., Wolf, Y., Sorokin, A., Mirkin, B., Koonin, E., Pavlov, A., Pavlova, N., Karamychev, V., Polouchine, N. et al. (2006)** Comparative genomics of the lactic acid bacteria. *Proc Natl Acad Sci U S A, 103, 15611-15616.*
17. **Molenaar, D., Bringel, F., Schuren, F.H., de Vos, W.M., Siezen, R.J. and Kleerebezem, M. (2005)** Exploring *Lactobacillus plantarum* genome diversity by using microarrays. *J Bacteriol, 187, 6119-6127.*
18. **Kurahashi, H., Inagaki, H., Yamada, K., Ohye, T., Taniguchi, M., Emanuel, B.S. and Toda, T. (2004)** Cruciform DNA structure underlies the etiology for palindrome-mediated human chromosomal translocations. *J Biol Chem, 279, 35377-35383.*
19. **Brodskii, L.I., Ivanov, V.V., Kalaidzidis Ia, L., Leontovich, A.M., Nikolaev, V.K., Feranchuk, S.I. and Drachev, V.A. (1995)** [GeneBee-NET: An Internet based server for biopolymer structure analysis]. *Biokhimiia, 60, 1221-1230.*
20. **Serrano, L.M., Molenaar, D., Wels, M., Teusink, B., Bron, P.A., de Vos, W.M. and Smid, E.J. (2007)** Thioredoxin reductase is a key factor in the oxidative stress response of *Lactobacillus plantarum* WCFS1. *Microb Cell Fact, 6, 29.*
21. **Storz, G., Opdyke, J.A. and Zhang, A. (2004)** Controlling mRNA stability and translation with small, noncoding RNAs. *Curr Opin Microbiol, 7, 140-144.*
22. **Horesh, Y., Amir, A., Michaeli, S. and Unger, R. (2003)** A rapid method for detection of putative

RNAi target genes in genomic data. *Bioinformatics, 19 Suppl 2, II73-II80.*

23. **Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E. and Mello, C.C. (1998)** Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature, 391, 806-811.*
24. **Kennerdell, J.R. and Carthew, R.W. (2000)** Heritable gene silencing in *Drosophila* using double-stranded RNA. *Nat Biotechnol, 18, 896-898.*
25. **Carmell, M.A., Zhang, L., Conklin, D.S., Hannon, G.J. and Rosenquist, T.A. (2003)** Germline transmission of RNAi in mice. *Nat Struct Biol, 10, 91-92.*
26. **Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I. and Koonin, E.V. (2006)** A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct, 1, 7.*
27. **Mojica, F.J., Diez-Villasenor, C., Soria, E. and Juez, G. (2000)** Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol, 36, 244-246.*
28. **Tang, T.H., Bachellerie, J.P., Rozhdestvensky, T., Bortolin, M.L., Huber, H., Drungowski, M., Elge, T., Brosius, J. and Huttenhofer, A. (2002)** Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc Natl Acad Sci U S A, 99, 7536-7541.*
29. **Tang, T.H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J.P. and Huttenhofer, A. (2005)** Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol Microbiol, 55, 469-481.*
30. **Jansen, R., Embden, J.D., Gaastra, W. and Schouls, L.M. (2002)** Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol, 43, 1565-1575.*
31. **Godde, J.S. and Bickerton, A. (2006)** The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol, 62, 718-729.*
32. **Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A. and Horvath, P. (2007)** CRISPR provides acquired resistance against viruses in prokaryotes. *Science,* **315**, 1709-1712.
33. **Sambrook, J., Fritsch, E.F., Maniatis, T. (1989)** Molecular Cloning: a Laboratory Manual. *Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, 2nd edition.*
34. **Ferain, T., Garmyn, D., Bernard, N., Hols, P. and Delcour, J. (1994)** *Lactobacillus plantarum ldhL* gene: overexpression and deletion. *J Bacteriol, 176, 596-601.*
35. **Edgar, R.C. (2004)** MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics, 5, 113.*
36. **Kuipers, O.P., Beerthuyzen, M.M., Siezen, R.J. and De Vos, W.M. (1993)** Characterization of the nisin gene cluster *nisABTCIPR* of *Lactococcus lactis*. Requirement of expression of the *nisA* and *nisI* genes for development of immunity. *Eur J Biochem, 216, 281-291.*
37. **Groot, M.N., Klaassens, E., de Vos, W.M., Delcour, J., Hols, P. and Kleerebezem, M. (2005)** Genome-based in silico detection of putative manganese transport systems in *Lactobacillus plantarum* and their genetic analysis. *Microbiology, 151, 1229-1238.*
38. **van Rooijen, R.J. and de Vos, W.M. (1990)** Molecular cloning, transcriptional analysis, and nucleotide sequence of *lacR*, a gene encoding the repressor of the lactose phosphotransferase system of *Lactococcus lactis*. *J Biol Chem, 265, 18499-18503.*
39. **Maniatis, T., Jeffrey, A. and Kleid, D.G. (1975)** Nucleotide sequence of the rightward operator of phage lambda. *Proc Natl Acad Sci U S A, 72, 1184-1188.*
40. **Bailey, T.L. and Gribskov, M. (1998)** Combining evidence using p-values: application to sequence homology searches. *Bioinformatics, 14, 48-54.*
41. **Durbin, R., Eddy, S., Krogh, A., Mitchison, G. (1998)** Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. *Cambridge: Cambridge University Press.*
42. **Zuker, M. (2003)** Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res, 31, 3406-3415.*
43. **Barrett, M.T., Scheffer, A., Ben-Dor, A., Sampas, N., Lipson, D., Kincaid, R., Tsang, P., Curry, B., Baird, K., Meltzer, P.S. et al. (2004)** Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc Natl Acad Sci U S A, 101, 17765-17770.*
44. **Hanahan, D. (1983)** Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol, 166, 557-580.*

45. **Johansson, M.L., Molin, G., Jeppsson, B., Nobaek, S., Ahrne, S. and Bengmark, S. (1993)** Administration of different *Lactobacillus* strains in fermented oatmeal soup: in vivo colonization of human intestinal mucosa and effect on the indigenous flora. *Appl Environ Microbiol, 59, 15-20.*

46. **Kerkhoven, R., van Enckevort, F.H., Boekhorst, J., Molenaar, D. and Siezen, R.J. (2004)** Visualization for genomics: the Microbial Genome Viewer. *Bioinformatics, 20, 1812-1814.*

47. **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.*

CHAPTER 5

T-box regulated genes and T-box evolution in prokaryotes; an *in silico* analysis.

Michiel Wels
Tom Groot Kormelink
Michiel Kleerebezem
Roland J. Siezen
Christof Francke

Submitted for publication

Acknowledgements
We would like to thank Victor de Jager for his assistance in applying the RFAM calculations, Jos Boekhorst and Richard Notebaart for insightful discussions. Christof Francke gratefully acknowledges the support of the Netherlands Genomics Initiative via the Kluyver Centre for Genomics of Industrial Fermentations and the NBIC/BioRange program. For our work we used the life-science grid, which is part of the Dutch e-Science Grid (BIG GRID; www.biggrid.nl) and which is a collaborative effort of Netherlands National Computing Facilities foundation (NCF), the Netherlands Bioinformatics Centre (NBIC; www.nbic.nl) and the National Institute for Nuclear Physics and High Energy Physics (NIKHEF).

All supplementary files regarding this chapter can be found at www.cmbi.ru.nl/~mwels/Chapter_5

Phylogenetic footprinting, motif searching and RNA structure prediction procedures were employed to identify T-box elements and their specifier codons in all currently sequenced prokaryotic genomes. Despite their supposed ancient occurrence in bacteria the taxonomic distribution of T-boxes appeared remarkably restricted. T-box elements are abundantly present in Firmicutes species, with an average of sixteen representatives per genome, whereas only ten species outside the phyla of *Firmicutes* and *Actinobacteria* were found to contain T-boxes. Our findings confirmed the current view that the vast majority of the T-box regulated genes (>92%) is directly related to charging uncharged tRNA (i.e. via tRNA ligation, amino acid biosynthesis and amino acid transport). The specifier codon of the various T-box elements was used to improve the functional annotation of approximately 125 genes, including many genes that are notoriously difficult to annotate on basis of sequence similarity like amino acid transporters. A detailed analysis of especially this group of T-box elements indicated that the T-box should behave as an independently evolving functional module and can easily switch amino acid specificity. We hypothesize that the ancestral T-box was linked to one of the genes encoding a tRNA ligase of a branched chain amino acid, probably Ile.

Introduction

Transcription anti-termination is a regulatory mechanism commonly encountered in all lineages within the bacterial kingdom (for a review, see (1,2)). In transcription anti-termination, the regulation of transcription occurs after the initiation of RNA synthesis, but before transcription of the coding region. The mechanism of anti-termination involves a structural change in the RNA transcript that is dependent on the interaction of the transcript with either a regulatory protein (3), a tRNA (4) or a metabolite (5). The structural elements that compose these anti-terminators are correlated to conserved motifs on the DNA and can therefore be found by conserved sequence motif searches in upstream regions of regulated genes (6).

A well-studied anti-termination element is the so-called T-box. T-box anti-termination is an elegant and sensitive mechanism by which many bacteria maintain constant levels of tRNA charged with amino acids (7). In case of sufficient supply of charged tRNA in a cell, the T-box folds into a terminator structure, thereby blocking further transcription. Transcription can only proceed upon conversion into an anti-terminator structure, which is induced by binding of uncharged tRNA (8). Although anti-terminator formation involves contacts between many nucleotides, the specificity of the interaction is largely dependent on the interaction of a tri-nucleotide in the so-called specifier loop of the T-box with the anti-codon of an amino acid-specific tRNA (9). The appropriate assignment of the specifier codon has been used previously to improve the functional annotation of various gene(s) located downstream of the T-box (7,10-13). The T-box controlled genes identified thus far encode functionalities that reflect perfectly the pivotal role of uncharged tRNAs in the regulatory mechanism. These include not only tRNA ligation, but also amino acid biosynthesis and transport (7). The encoded proteins are involved in modulation of the level of uncharged tRNA in the cell, either directly by charging the corresponding tRNA with its cognate amino acid and/or indirectly by controlling the intracellular concentration of the specific amino acid. To date, T-boxes have predominantly been identified in the genomes of bacterial species of the phyla *Firmicutes* (including *Mollicutes*) and *Actinobacteria* (6). However, anti-termination systems are considered to be among the oldest regulatory systems in bacteria because of their independence of regulatory proteins and occurrence in different phyla (14). To investigate this further, we have explored the occurrence of T-boxes in all sequenced prokaryotic genomes to obtain the (almost) full complement of this regulatory element. To circumvent potential differences between T-box systems in different bacterial lineages, an iterative HMM-based identification method was applied using the best conserved region of the T-box sequence. Species- and amino acid-specific T-box regulation networks were reconstructed. The acquired knowledge on amino acid specificity could be used to propose an improved functional annotation for many T-box controlled genes and to shed light on the evolution of the regulatory element itself.

Materials and Methods

Sequence information and tools

Genome sequences and annotation files of completely sequenced bacterial and archaeal genomes were downloaded from the NCBI repository (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) on November 1st 2005. In addition, for the analysis of the molecular functions encoded by T-box controlled genes, genomic information was obtained from the ERGO genome analysis and discovery system (15) and updated until March 1st 2007. Conserved nucleotide sequence motifs were recovered via MEME (16) (settings: maximally 5 motifs, modus ZOOPs, minimal width 10 nt, maximal width 30 nt) and visualized using Weblogo (17). Multiple sequence alignments were created with MUSCLE (18) and bootstrapped Neighbor-Joining trees with CLUSTALX (19) (corrected for multiple substitutions). Hidden Markov Models (HMMs) were constructed from a multiple sequence alignment using HMMER 2.3.2 (20). RNA secondary structure predictions were performed by Mfold (21) using standard settings.

Identification of a general T-box sequence

The T-boxes reported so far in literature were mainly located upstream of tRNA ligase encoding genes in species of the phylum *Firmicutes* (6,22). We therefore initiated the search for a general T-box sequence motif by collecting the nucleotide sequences (300 nt) preceding all tRNA ligases (n=910) found in the sequenced genomes of the *Firmicutes*. The nucleotide sequences were searched for conserved motifs, and four characteristic consecutive motifs were recovered. The four identified motifs are part of a generic T-box sequence that spans a length of about 250 nt, which is comparable to the mean size of T-boxes observed by (7). The best-conserved motif (E-value=$1.2e^{-2841}$), which is located closest to the translation start, had a length of 30 nt and was found in about 61% (n=553) of the nucleotide sequences. These 553 sequences of 30 nt were aligned and an HMM was constructed to search the database of sequenced genomes for the motif. This search yielded 374 new hits, making a total of 927 putative T-boxes. The upstream 500 nt of every hit was then checked for the presence of the three other T-box specific motifs recovered using MEME. In only 40 cases none of the other motifs could be detected and these hits were therefore considered false-positives and were hence removed. On basis of this result we estimated the false-positive rate of the HMM detection at ~4%. 27 of the 40 apparently false positive hits were solitary (i.e. one hit per organism) and, moreover, not situated in (the correct direction of) the proximal upstream region of a gene. The remaining 887 T-box sequences were subdivided on basis of the taxonomy of the species. Class-specific HMMs were built from the sequences retrieved from the best MEME motif, together with 10 nt upstream and 5 nt downstream of the motif. A search was performed with the new HMMs and in case new hits were found, a new alignment and HMM were generated and the search was repeated until no new sequences were retrieved. This iteration procedure yielded only 10 additional T-boxes. These T-boxes were manually checked for the presence of the other three specific motifs. The low number of additional hits implies an extremely low false-negative rate (Table 1 gives the distribution of the T-boxes over the different phyla).

Functional classification of the genes regulated by T-boxes

The genes downstream of the recovered T-boxes were divided into four different classes on basis of the gene annotation information: ‘tRNA ligation ’, ‘amino acid biosynthesis, ‘amino acid transport’ and ‘other’. Operon structures, defined as runs of genes following the T-box located on the same DNA strand with an intergenic region smaller than 100 bp, were taken into account in the analysis. If differences in operon lengths were found between operons of species of the same taxonomic family (e.g. *Bacilli*), the genome organization was manually inspected for missed or over-predicted genes. Errors were corrected when necessary.

Comparison with the RFAM T-box model

The T-box HMM from the RFAM database (23) was used to scan a selection of genomes for T-boxes. The set was chosen such that it included at least one representative for each phylum and at least one representative of the major orders within the phylum *Firmicutes* (*Bacillales, Clostridia, Lactobacillales* and *Mollicutes)*. The set included well-studied organisms like *Escherichia coli, Yersinia pestis, Bacillus subtilis, Listeria monocytogenes, Staphylococcus aureus, Lactococcus lactis* and *Clostridium acetobutylicum*. When using a cut-off of 53.000 bits (described by RFAM as reliable) RFAM predicted only half of the T-boxes identified by our method. As almost all the additional T-boxes we identified were found in the upstream regions of genes encoding tRNA ligases or proteins related to amino acid biosynthesis and transport, our extra hits seem valid. We observed that many of those were also found using the RFAM model in case a lower cutoff-value was chosen. However at the same time, the decrease in cutoff-value caused a rapid increase in the number of false-positives using the RFAM model; 51% of the new hits were located within the coding regions of genes that do not seem to be related to tRNA ligation or amino acid metabolism. Exemplary for this finding was the identification of over a hundred additional T-boxes within the genomes of *Clostridium acetobutylicum, Clostridium tetani* and *Mycoplasma mycoides*. The scan was limited to a representative set of species because of the fact that the RFAM-search was extremely computationally expensive (more than two weeks on a 8 node (16 core) linux cluster). In addition to predicting far more T-boxes without an increase in false-positives, our new method was far less computationally intensive.

Results

A comprehensive collection of T-boxes

The analysis of the taxonomic and functional distribution of T-boxes was started by *de novo* identification of T-box motif characteristics. Conserved nucleotide sequence motifs upstream of tRNA ligase encoding genes in species of the phylum *Firmicutes* were searched and used to identify T-boxes located at other positions in the same genome as well as in the genomes of other species (see methods). These searches showed that a T-box could be specified best by a 30 nt motif that is extremely well-conserved and positioned in the 3'-region of the terminator/anti-terminator loop (Supplementary Figure S1). In fact, this motif is known as 'the T-box sequence' since its discovery (24). Later it was recognized that this conserved region belongs to a larger conserved RNA structure known as the T-box element (8).

We observed prominent variations in the number of T-boxes per genome between different classes of the phylum *Firmicutes* and between different phyla. Therefore, additional searches with phylum/class-specific HMMs were performed. The phylum-specific and class-specific T-box HMMs generally did not yield novel hits, except in the case of the *Clostridia* where 10 new T-boxes were identified. Additional iterations did not expand the dataset further. Considering the high level of conservation of 'the T-box sequence' in all recovered T-boxes and the correlated presence with other conserved T-box regions, the above implied that the recovered collection of T-boxes should be (virtually) complete for all species investigated. We compared the number of T-boxes identified by us for a representative set of organisms with the number obtained using the RFAM T-box model (23) and found that our method was superior in all aspects (see methods). Firstly, we recovered twice the amount of T-boxes using the recommended settings for the RFAM model and our collection included all those recovered by the RFAM model. Secondly, the computational capacity required for acquiring the data was 1000-fold less when employing our method.

T-box regulated genes

As expected, the proposed regulatory role of the T-box elements appeared to be perfectly reflected by the genes under their control. The majority of the T-boxes (62%) were found to precede genes encoding tRNA ligases, while most others were found upstream of genes encoding proteins involved in amino acid transport (12%) or amino acid biosynthesis (18%). The remaining T-boxes (8%; 71 genes in total) were found upstream of genes encoding proteins with unknown function (54 genes), or a function that lacks an apparent relation to amino acid metabolism (17 genes). A complete and species-specific subdivision of T-boxes based on function prediction of the proteins encoded downstream and a list of genes with no apparent relation to amino acid metabolism is provided in the supplementary material, Tables S2 and S3.

Taxonomic variations in the distribution of T-box regulation

We generated a comprehensive list of T-boxes that are present in all sequenced genomes (see Table 1 for the phylogenetic distribution) and confirmed the previous finding that T-boxes are exclusively present in bacterial genomes and are predominantly encountered in species of the phylum *Firmicutes* (>95% of the hits) (22). Moreover, our analyses uncovered many previously unidentified T-boxes in the genomes of Firmicutes as well as in species from other phyla. In species of the class *Mollicutes* two T-box elements were found in two members (constituting the subclass *Endoplasmatales* (25)) out of twelve for which the genome has been completely sequenced. In *Mycoplasma mycoides* and *Mesoplasma florum* a T-box is located in front of two tRNA ligase encoding genes (for isoleucine: *ileS*; and for threonine: *thrS*). The finding of a characteristic and conserved regulatory element in *Mycoplasma mycoides* and *Mesoplasma florum* but not in other *Mollicutes,* strengthens the current taxonomic division.

Table 1: The occurrence of T-boxes in different bacterial phyla.
The phyla are taken from the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/).

Phylum	Genomes sequenced	Genomes with at least one T-box	Number of T-boxes
Firmicutes	53	53	855
Actinobacteria	19	12	32
Chloroflexi	2	2	6
Deinococcus/Thermus	3	1	2
Proteobacteria	125	2	2
Cyanobacteria	13	0	0
Chlamydiae	10	0	0
Bacteroidetes/Chlorobi	7	0	0
Spirochaetes	6	0	0
Planctomycetes	1	0	0
Aquificiates	1	0	0
Fusobacteria/Thermotogae	2	0	0
Total	242	70	897

In *Proteobacteria* (i.e. in *Geobacter sulfurreducens* and *Pelobacter carbinolicus*) two typical T-box elements were identified upstream of the *leuA* gene (encoding a 2-isopropylmalate synthase, EC: 2.3.3.13). Also some species of the phyla *Deinococcus* and *Chloroflexi* contain T-box elements, but these are not conserved throughout those phyla. In the *Deinococcus/Thermus* group, T-boxes were found in *Deinococcus radiodurans* but not in any of the *Thermus sp.* genomes. In *D. radiodurans* the T-boxes are located in the upstream regions of *ileS* and *glyS* (tRNA ligase for glycine). In the *Chloroflexi* genomes of *Dehalococcoides ethenogenes* and *Dehalococcoides CBDB* three T-boxes were identified upstream of *ileS* and upstream of two operons involved in tryptophan biosynthesis, respectively. Members of the phyla *Cyanobacteria, Chlorobi* and *Spirochaetes* appeared to lack T-box elements.

Earlier analysis of riboswitches in Actinobacteria showed that some species belonging to this phylum contain a T-box upstream of *ileS* (10). We found that *Symbiobacterium thermophilum*, although currently classified as an *Actinobacterium* on basis of rRNA sequences, contains not less than eighteen T-boxes. This number is comparable to the numbers found in species of the *Firmicutes*. This finding is in line with a previous study that concluded that *S. thermophilum* is probably more closely related to *Firmicutes* than to *Actinobacteria* (26). Therefore, *S. thermophilum* will be treated as a member of the *Firmicutes* phylum in this paper.

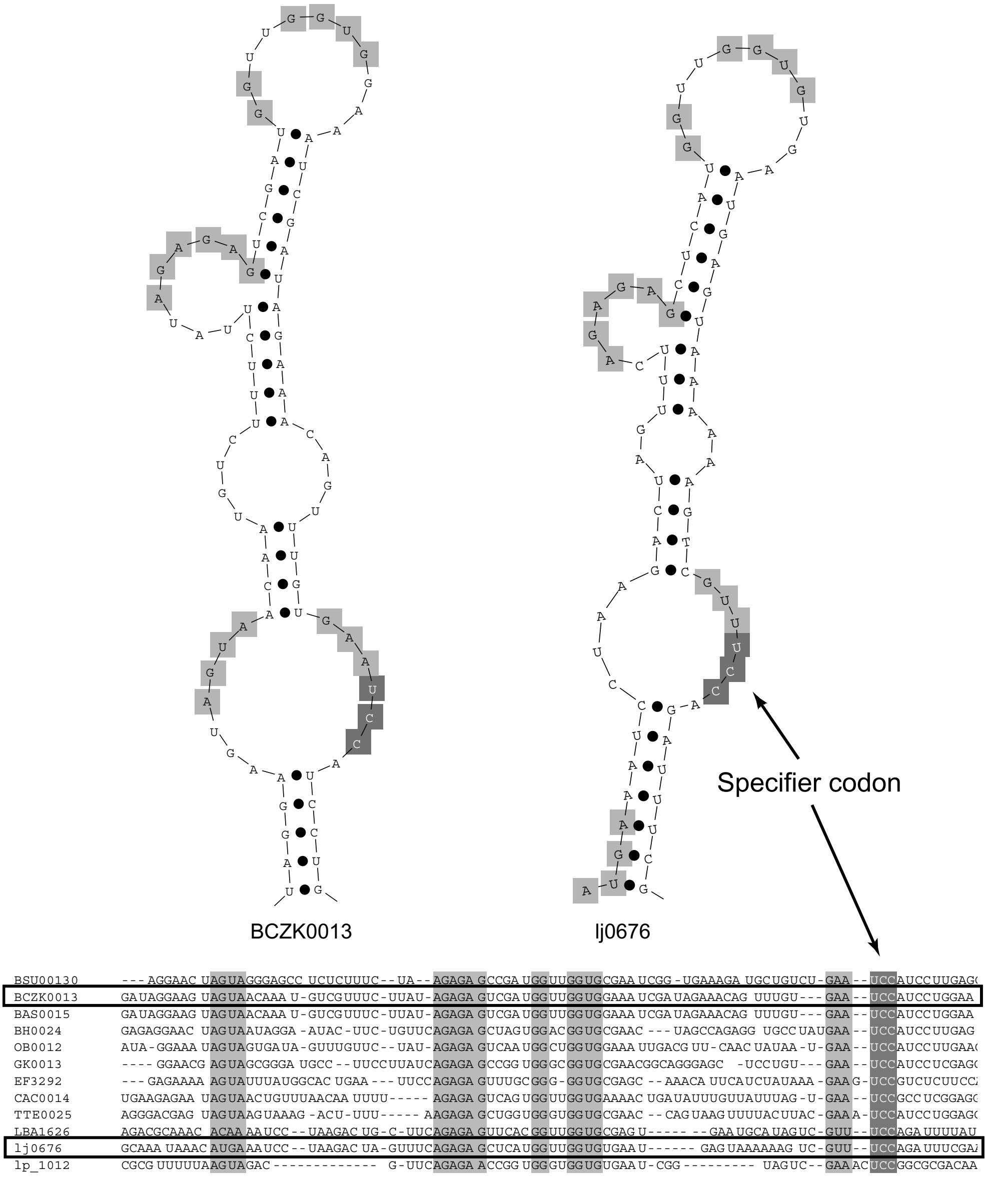
Identification of the specifier codon and amino acid specificity.

Although T-boxes were readily identified, it was more difficult to define their amino acid specificity. To that end, homologous proteins (detected using the Smith Waterman algorithm, E-value <1 e-10) from different species encoded by genes preceded by a T-box were clustered in a Neighbor-Joining tree and the upstream regions corresponding to each cluster in the tree were aligned (the alignments can be found at http://www.cmbi.ru.nl/T_box_analysis). Specifier codons were assigned by combining the information on sequence conservation of the aligned upstream regions with a secondary structure prediction of the specifier-loop containing part of the T-box. For specifier loop structure prediction, a MEME-generated motif containing the best conserved region of the specifier loop (containing the conserved AGAGA and GAA box) was used. A sequence containing this motif together with an extra 30 bp. up- and downstream was used as input for the structure prediction (see methods). In the vast majority of cases the specifier codon was already apparent from the alignment, whereas in some other cases, the secondary structure prediction provided the additional information required to define the specifier codon in the specifier loop. The combined analysis of alignment and secondary structure allowed codon prediction for more than 90% of the identified T-boxes. Figure 1 depicts an example of the combination of a multiple sequence alignment and secondary structure prediction for the specifier-loop containing part of the T-boxes preceding the genes encoding Ser-tRNA ligase in different species. Remarkably, inspection of the various secondary structure predictions showed that the well-conserved nucleotides within the various T-boxes are located mainly inside the loops and not within the parts that are predicted to form double strands.

To investigate the variation in T-box sequence further a multiple sequence alignment and a bootstrapped Neighbor Joining (NJ) tree was generated for the complete set (not shown). If amino acid specificity would be dominant for sequence conservation then this should be reflected in such a tree by clustering of the T-boxes on basis of the amino acid specificity. However, the NJ tree had an extremely shattered appearance with a major scattering of identified amino acid specificity of T-boxes over many clusters in the tree. Moreover, the bootstrap support for the vast majority of clusters was low. It therefore appeared that the similarity between the amino acid specific T-boxes was rather low. In addition, a comparison of the weblogos created on basis of the amino acid specific alignments did not reveal characteristic differences outside the nucleotides determining the specifier codon (see additional files at http://www.cmbi.ru.nl/T_box_analysis). The consequences of these findings will be discussed later in this paper (see discussion).

Improved annotation.

In all cases, the T-boxes located upstream of the genes encoding a tRNA ligase contained a specifier codon that corresponded with the amino acid specificity of the ligase. Furthermore, in all other cases where the function of the protein encoded by the gene downstream of the T-box had experimentally been verified, the specifier-codon corresponded to the established functionality of the gene. The alignments and the annotation of the different specifier codons can be found at http://www.cmbi.ru.nl/T_box_analysis. The above implies that the employed method for the identification of T-box specificity is reliable and, consequently, that predicted T-box specificities can be extrapolated to the molecular function of the protein encoded by the gene located downstream, as has occasionally been done before. This approach proved very useful as many of the genes preceded by a T-box had not been specifically annotated to date. In fact, more than two-third of the non-tRNA ligase genes preceded by a T-box lacked a specific annotation.



```
BSU00130    ---AGGAAC UAGUAGGGAGCC UCUCUUUC--UA--AGAGA GCCGA UGGUUGGUGCGAA UCGG-UGAAAGA UGCUGUCU-GAA--UCCAUCCUUGAG
BCZK0013    GAUAGGAAG UAGUAACAAA U-GUCGUUUC-UUAU-AGAGA GUCGA UGGUUGGUGGAAA UCGA UAGAAACAG  UUUGU---GAA--UCCAUCCUGGAA
BAS0015     GAUAGGAAG UAGUAACAAA U-GUCGUUUC-UUAU-AGAGA GUCGA UGGUUGGUGGAAA UCGA UAGAAACAG  UUUGU---GAA--UCCAUCCUGGAA
BH0024      GAGAGGAAC UAGUAAUAGGA-AUAC-UUC-UGUUCAGAGA GCUAGUGGACGGUGCGAAC ---UAGCCAGAGG  UGCC UAUGAA--UCCAUCCUUGAG
OB0012      AUA-GGAAA UAGUAGUGAUA-GUUUGUUC--UAU-AGAGA GUCAA UGGCUGGUGGAAA UUGACG UU-CAAC UAUAA-U-GAA--UCCAUCCUUGAA
GK0013      ----GGAACG AGUAGCGGGA UGCC--UUCCUUAUCAGAGA GCCGG UGGGCGGUGCGAACGGCAGGGAGC  -UCCUGU---GAA--UCCAUCCUCGAG
EF3292      ---GAGAAAA AGUAUUUAUGGCAC UGAA---UUCCAGAGA GUUUGCGGG-GGUGCGAGC --AAACA UUCAUCUAUAAA-GAAG-UCCGUCUCUUCC
CAC0014     UGAAGAGAA UAGUAACUGUUUAACAA UUUU----- AGAGA GUCAG UGGUUGGUGAAAAC UGAUAUUUGUUAUUUAG-U-GAA--UCCGCC UCGGAG
TTE0025     AGGGACGAG UAGUAAGUAAAG-ACU-UUU----- AAGAGA GCUGGUGGGUGGUGCGAAC --CAGUAAGUUUUACUUAC-GAAA-UCCAUCCUGGAG
LBA1626     AGACGCAAAC ACAA AAUCC--UAAGAC UG-C-UUCAGAGA GUUCAC GGUUGGUGCGAG U----- --GAA UGCA UAGUC-GUU--UCCAGA UUUUAU
lj0676      GCAAA UAAAC AUGA AAUCC--UAAGAC UA-GUUUCAGAGA GCUCA UGGUUGGUGUGAA U------ GAGUAAAAAAG UC-GUU--UCCAGA UUUCGA
lp_1012     CGCG UUUUUAAGUAGAC-------------  G-UUCAGAGA ACCGG UGGGUGGUGUGAA U-CGG--------- UAGUC--GAAACUCCGGCGCGACAA
```

Figure 1: The identification of the specifier-codon using multiple sequence alignment and secondary structure prediction. (Bottom) A multiple sequence alignment of the specifier-loop containing part of Ser tRNA-ligase related T-boxes. Conserved residues are shown in shaded boxes. (Top) Secondary structure prediction for the specifier-loop containing part of the T-box preceding the gene encoding Ser tRNA-ligase in Bacillus cereus ZK and Lactobacillus johnsonii *(BCZK0013* and *ljo0676*, respectively). Although the overall sequence conservation was low for this part of the T-box, the predicted overall secondary structure appeared very similar and the specifier loop could easily be identified. The sequence conservation in the loop was used to define the specifier-codon. In this case the codon was UCC, which corresponds to serine.

T-box regulation of tRNA ligases in Firmicutes.

To evaluate the functional distribution of T-boxes in more detail, the correlation between tRNA ligases and T-box regulatory elements was analyzed in depth for the Firmicutes (Figure 2). It appears that regulation of T-boxes is conserved in almost all (at least 29 out of 34) *Firmicutes* for several tRNA ligases (*ileS*, a*laS*, *serS* and *thrS*), whereas some tRNA ligases (*lysS, asnS* and *gltX*; the latter gene encodes a Glu-tRNA ligase (see also the legend of Figure 2)) appeared to be controlled by a T-box in only a few species. The genes encoding the tRNA ligases for cysteine and asparagine were often found as the second gene in a putative operon that was T-box regulated. In those cases of shared T-boxes, the gene encoding the Asp-tRNA ligase was always preceded by a gene encoding a His-tRNA ligase, while the gene encoding the Cys-tRNA ligase was always preceded by a gene encoding a serine acetyltransferase. The latter protein catalyzes one of the intermediate steps in the biosynthesis of Cysteine. In several organisms, multiple copies of amino acid specific tRNA ligase encoding genes are found and in more than half of the cases (58%) only one of them is subject to T-box regulation.

A clear phylogenetic effect is observed when the four major orders within the *Firmicutes* (*Bacillales*, *Clostridia*, *Lactobacillales* and *Mollicutes*) are compared (Figure 2). This is true for both the number and the type of tRNA ligase encoding genes regulated by a T-box. Most T-box regulated tRNA ligase encoding genes are found in species of the *Bacillus anthracis/cereus* group within *Bacillales*. The only tRNA ligase encoding gene that is not regulated by a T-box in at least one of the sequenced *cereus* group genomes is *gltX*. In other organisms of the *Bacillales* order (including *Staphylococcus* and *Listeria* species) the number of T-box regulated tRNA ligase encoding genes is lower, but still relatively high (>50%).

Within the *Lactobacillales*, it appears that *Streptococci* have far less tRNA ligase encoding genes regulated by T-boxes than *Lactobacillus* species. The lowest number of tRNA ligases regulated by T-boxes is found for *S. thermophilus* and *S. pneumoniae*. The relatively low amount of T-box regulation in these species could be the result of regressive evolution, a process that was suggested to be the underlying mechanism for the large loss of functionally active genes in *S. thermophilus* (27). Analogously, the pathogenic *Mollicutes* lack all-but-two T-box elements. Another interesting observation is the diversity in the regulation by T-boxes of the tRNA ligase encoding genes in different *S. pyogenes* strains (Figure 2). As the different T-boxes are found across the complete lineage of *Firmicutes* and lack only in a few *Streptococci*, this strain-specific distribution is probably the result of recent T-box element loss in these species and strains, in line with the suggested regressive evolution.

T-box regulation of genes involved in amino acid biosynthesis in the Firmicutes

T-box regulation of genes related to amino acid biosynthesis has been described previously for various amino acids (12,13,28). Our generic analysis confirms the importance of T-box control in the biosynthesis of aromatic amino acids and methionine (Figure 3). In those cases were the regulation was studied before, our analysis confirmed earlier studies (12,13,28). In concordance with previous observations made for *B. subtilis* (29) we found that the conversion of serine into cysteine, as well as the synthesis of branched chain amino acids is T-box controlled in a variety of related organisms (Figure 3). In fact, we observed that (part of the) tryptophan, cysteine, leucine,

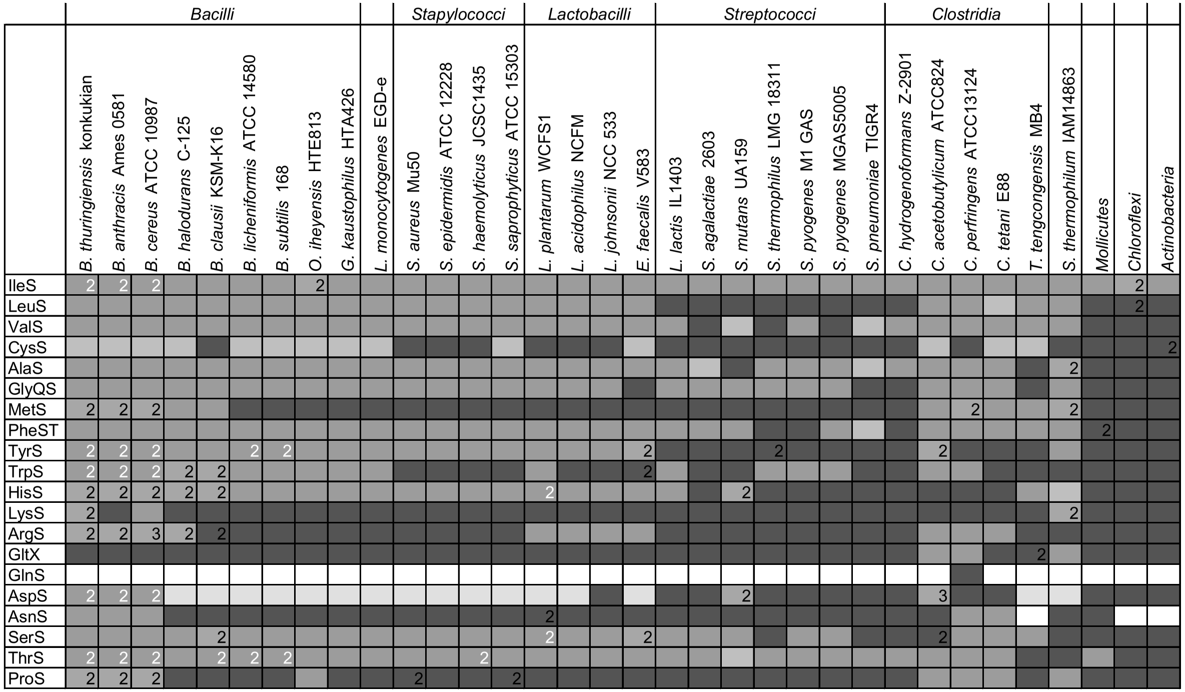

Bacilli
Stapylococci
Lactobacilli
Streptococci
Clostridia
B. thuringiensis konkukian
B. anthracis Ames 0581
B. cereus ATCC 10987
B. halodurans C-125
B. clausii KSM-K16
B. licheniformis ATCC 14580
B. subtilis 168
O. iheyensis HTE813
G. kaustophilus HTA426
L. monocytogenes EGD-e
S. aureus Mu50
S. epidermidis ATCC 12228
S. haemolyticus JCSC1435
S. saprophyticus ATCC 15303
L. plantarum WCFS1
L. acidophilus NCFM
L. johnsonii NCC 533
E. faecalis V583
L. lactis IL1403
S. agalactiae 2603
S. mutans UA159
S. thermophilus LMG 18311
S. pyogenes M1 GAS
S. pyogenes MGAS5005
S. pneumoniae TIGR4
C. hydrogenoformans Z-2901
C. acetobutylicum ATCC824
C. perfringens ATCC13124
C. tetani E88
T. tengcongensis MB4
S. thermophilum IAM14863
Mollicutes
Chloroflexi
Actinobacteria
IleS
LeuS
ValS
CysS
AlaS
GlyQS
MetS
PheST
TyrS
TrpS
HisS
LysS
ArgS
GltX
GlnS
AspS
AsnS
SerS
ThrS
ProS

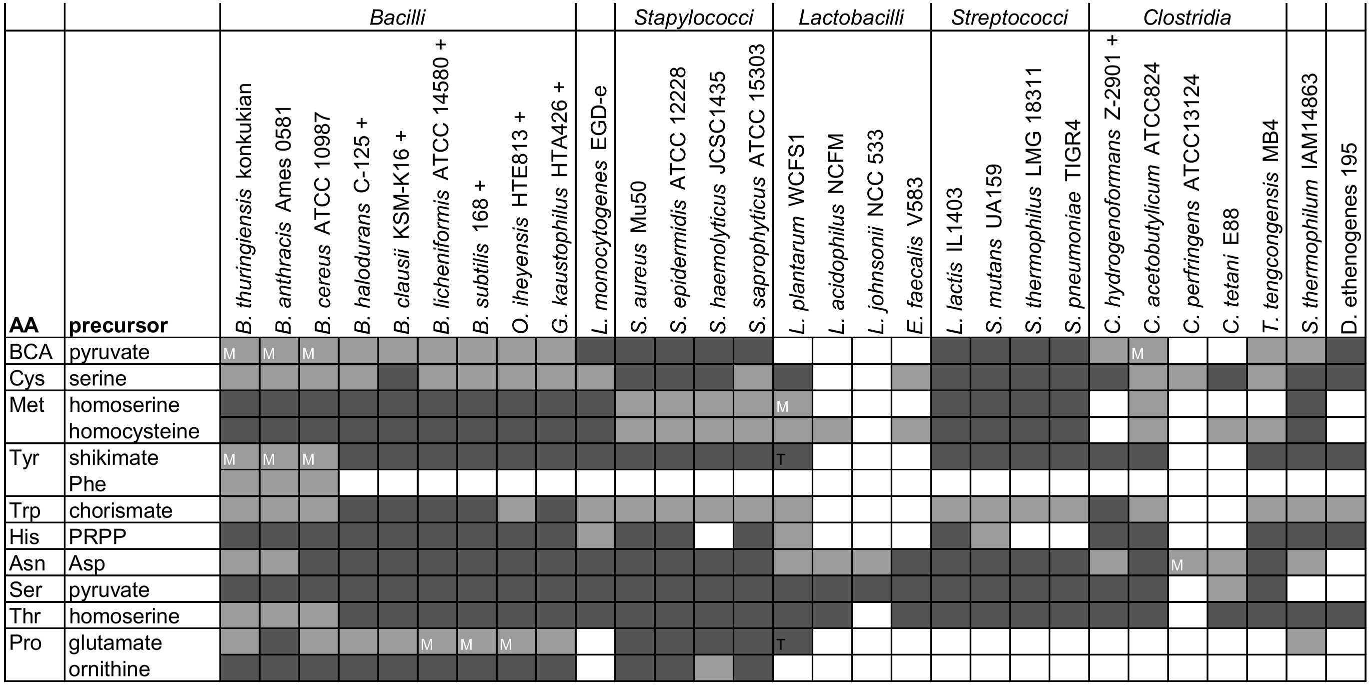
Figure 2 (a color version of the figure can be found at page 172): T-box regulation of tRNA ligase encoding genes in the *Firmicutes*.
The color coding relates to the presence or absence of a T-box upstream of the genes encoding the amino acid-specific tRNA ligases in the various species and strains. Green indicates the tRNA ligase(s) is (are) regulated by a T-box and red that the tRNA ligase(s) is (are) not regulated by a T-box. Although most tRNA ligases are present in one copy on the genome, several organisms contain two, or in some cases three copies of specific ligases (indicated by a number in the box). Orange indicates that 1 of the 2 tRNA ligases is regulated by a T-box or 1 out of 3 in the case of the *argS* genes in *B. cereus* ATCC 10987 and the *aspS* genes in *C. acetobutylicum*. Light green indicates that the tRNA ligase is not the first in the operon, but is regulated by a T-box with the same specificity. Yellow color coding indicates that the regulated tRNA ligase is the second gene in an operon in combination with another tRNA ligase gene. These genes are probably regulated by a T-box with different specificity than the specificity of the tRNA ligase. White indicates that no tRNA ligase of this type present in the organism. In principle, a species needs at least one specific tRNA ligase for each amino acid. Nevertheless, there are exceptions. For instance, all but one *(Clostridium perfringens)* of the analyzed genomes, lack the gene that encodes a Gln-tRNA ligase and the genomes of the *Chloroflexi, Actinobacteria* and *Thermoanaerobacter tencongens* also lack an Asn-tRNA ligase. In these cases, the biological role of the Gln-tRNA ligase is taken over by the Glu tRNA ligase, which couples a Glu residue to the tRNAGln. The residue is subsequently transformed into a Gln by a tRNA specific amidotransferase (50). Similarly, an Asn-tRNAAsn is formed via transamidation of an Asp residue (Asp-tRNAAsn to Asn-tRNAAsn) in bacteria that lack an Asn tRNA ligase (51). Consequently, we found that all species lacking either the Gln-tRNA ligase or the Asn-tRNA ligase have an orthologous gene coding for the corresponding amidotransferase. No T-boxes were identified upstream of those genes.

isoleucine, valine, asparagine, proline, tyrosine, threonine, methionine, histidine and serine biosynthesis routes are controlled by T-boxes, but that the distribution over the analyzed species is rather variable. For the biosynthesis of Branched Chain Amino Acids (BCA: isoleucine, leucine, valine) our data display a clear phylogenetic pattern in which all members within a taxonomic family either do, or do not regulate BCA synthesis by a T-box. It is found in members of the order of *Bacillales* and the order of *Clostridia*, although several families within the order of *Bacillales* (e.g. *Staphylococci* and *Listeria*) as well as several *Streptococci* consistently lack T-box control of BCA biosynthesis. Expression of BCA biosynthesis genes was previously shown to be controlled by both T-box as well as different regulatory proteins (*CodY* and *TnrA*) in *B. subtilis* (30-32). For tyrosine biosynthesis, T-boxes were also found to be phylogenetically conserved. In addition to a previous study that showed that tyrosine biosynthesis from shikimate is T-box regulated in *B. anthracis* (13), we found that all bacteria belonging to the *B. cereus* group (of which three are shown in figure 3) contain a T-box in the upstream region of the related operon. In addition, the *B. cereus* group representatives are the only organisms in our study that encode a phenylalanine-4-hydroxylase ortholog (converting phenylalanine into tyrosine). We found this gene also to be T-box regulated in all members of the *B. cereus* group.

T-box regulation of amino acid transport.

Many genes encoding amino acid transporters were found preceded by a T-box, especially in the genomes of the *Lactobacilli* and the *Bacilli* of the *Bacillus cereus*-group. The identified T-box controlled transporters belonged to no less than seven distinct transporter families (MFS, 2.A.1; APC, 2.A.3; NSS, 2.A.22; DAACS, 2.A.23; LIVCS, 2.A.26; NhaC, 2.A.35; and ABC-cassette, 3.A.1). These families are related to the two main transporter sub-classes “porters” and “P-P bond hydrolysis driven”, according to the Transporter Classification described by Saier (33). The first class of transport systems uses membrane potential (via cation symport and solute antiport) to acquire solutes, whereas the second class of transport systems apply the free-energy obtained by the hydrolysis of ATP. Table 2 gives an overview of the distribution of the transport systems regulated by a T-box over the various Firmicutes species.

The identification of the amino acid specificity of the T-box enabled the improvement of the functional annotation of downstream transporters in the vast majority of cases. Overall, for more than 85% of the T-box regulated transporters the specificity could be improved as compared to the entries in the reference database of NCBI. A full list can be found in the supplementary material (Table S2). We have limited the substrate specificity in our annotation to putatively dominant substrates. However, broader substrate specificity is probably more common for transporters. Especially transport systems consisting of only a permease are expected (and have been shown) to display broader substrate specificity (see (34) and (35) for examples). In all cases the amino acid specificity of the T-box proved valuable in assessing the biological role of the controlled systems. We will discuss the T-box based identification of some transport systems in more detail in the following paragraphs and in additional file 1 for the T-box regulated ABC transport systems.



		Bacilli										Stapylococci				Lactobacilli				Streptococci				Clostridia						
AA	**precursor**	*B. thuringiensis* konkukian	*B. anthracis* Ames 0581	*B. cereus* ATCC 10987	*B. halodurans* C-125 +	*B. clausii* KSM-K16 +	*B. licheniformis* ATCC 14580 +	*B. subtilis* 168 +	*O. iheyensis* HTE813 +	*G. kaustophilus* HTA426 +	*L. monocytogenes* EGD-e	*S. aureus* Mu50	*S. epidermidis* ATCC 12228	*S. haemolyticus* JCSC1435	*S. saprophyticus* ATCC 15303	*L. plantarum* WCFS1	*L. acidophilus* NCFM	*L. johnsonii* NCC 533	*E. faecalis* V583	*L. lactis* IL1403	*S. mutans* UA159	*S. thermophilus* LMG 18311	*S. pneumoniae* TIGR4	*C. hydrogenoformans* Z-2901 +	*C. acetobutylicum* ATCC824	*C. perfringens* ATCC13124	*C. tetani* E88	*T. tengcongensis* MB4	*S. thermophilum* IAM14863	D. ethenogenes 195
BCA	pyruvate	▨ M	▨ M	▨ M	▨	▨	▨	▨	▨	▨	■	■	■	■	■					■	■	■	■	▨	▨ M			▨	▨	■
Cys	serine	▨	▨	▨	▨	■	▨	▨	▨	▨	▨	■	■	■	▨	■			▨	■	■	■	■	■	▨	▨	■	▨	■	■
Met	homoserine	■	■	■	■	■	■	■	■	■	■	▨	▨	▨	▨	▨ M				■	■	■	■		▨				■	
	homocysteine	■	■	■	■	■	■	■	■	■	■	▨	▨	▨	▨	▨	▨		▨	■	■	■	■		▨		▨	▨	■	
Tyr	shikimate	▨ M	▨ M	▨ M	■	■	■	■	■	■	■	■	■	■	■	■ T				■	■	■	■	■	■			■	■	■
	Phe	▨	▨	▨																										
Trp	chorismate	▨	▨	▨	■	■	■	■	▨	■	▨	▨	▨	▨	▨	▨				▨	▨	▨	▨	■	▨			▨	▨	▨
His	PRPP	■	■	■	■	■	■	■	■	■	▨	■	■		■	▨				■	▨			■	■			■	■	■
Asn	Asp	▨	▨	■	■	■	■	■	■	■	■	■	■	■	■	▨	▨	▨	■	■	■	■	■	▨	■	▨ M	▨	■	▨	
Ser	pyruvate	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		▨	■		
Thr	homoserine	▨	▨	▨	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■		■	■	■	■
Pro	glutamate	▨	■	▨	▨	▨	▨ M	▨ M	▨ M	▨		■	■	■	■	■ T													▨	
	ornithine	■	■	■	■	■	■	■	■	■		■	■	▨	■															

Figure 3 (a color version of the figure can be found at page 173): T-boxes preceding the genes related to amino acid biosynthesis in *Firmicutes.*
Color coding identifies the presence of the biosynthesis pathway and whether it is regulated by a T-box: green; T-box regulated; red; not T-box regulated; no color; pathway absent. +TRAP protein is present. M pathway genes organized in multiple operons. BCA indicates the branched chain amino acids valine, leucine and isoleucine.

Table 2: Overview of T-box regulated transporter genes in different Firmicutes.
The number and type of transporters are displayed per species and according to their predicted specificity. At the bottom the total fraction of regulated transporters is shown per family.

	B. anthracis Ames 0581	*B. licheniformis* ATCC 14580	*B. subtilis* 168	*O. iheyensis* HTE813	*L. monocytogenes* EGD-e	*L. plantarum* WCFS1	*L. acidophilus* NCFM	*L. johnsonii* NCC 533
Asn							ABC	
Asp								ABC
His					ABC	ABC	ABC	
Ile							ABC	ABC
	LIVCS					LIVCS	LIVCS	LIVCS
Leu		APC	APC					
	NSS							
Lys	MFS							
Met						ABC (5)	ABC (3)	ABC
Phe	NSS							
Thr	APC							
	LIVCS							
Trp						ABC		
	NSS							
Tyr	NHAC			NHAC		NHAC (2)	NHAC	NHAC
Val						LIVCS	LIVCS	
Unclear						ABC		
ABC					1|73	8|77	6|48	3|59
APC	1|19	1|20	1|18					
LIVCS	2|6					2|3	2|3	2|2
MFS	1|69							
NHAC	1|4			1|3		2|2	1|1	1|1
NSS	3|4							

E. faecalis V583	*L. Lactis* IL1403	*S. pneumoniae* R6	*C. acetobutylicum* ATCC824	*C. perfringens* ATCC13124	*C. tetani* E88
ABC					
			ABC		
			LIVCS	LIVCS	LIVCS
					LIVCS
ABC (3)	ABC				
		ABC			ABC
NHAC					
					NSS
				DAACS	
4|79	1|57	1|78	2|93		1|60
			1|1	1|3	2|4
1|2					
					1|5

The APC family

In the studied *Bacillus* genomes a T-box was identified in front of an APC-family protein encoding gene. In *B. subtilis* and *B. licheniformis* the gene *ybvW* is preceded by a Leu T-box, whereas such a box is lacking upstream of the orthologous genes, which are found in *E. faecalis*, *G. kaustophilus* and in *L. lactis* (co-orthologs: *yibG* and *ysjA*). The Leu T-box suggests that the YbvW protein is a Leucine transporter, in line with the general functionality of transporters of the APC family (family characteristics described in (36,37)). Surprisingly, in the members of the *Bacillus cereus*-group another APC family gene is preceded by a T-box, specific for threonine. Although an orthologous gene is present in most of the Firmicutes genomes, e.g. *ykbA* in *B. subtilis,* it is regulated by a T-box only in the species of the *Bacillus cereus*-group it is regulated by a T-box. When the upstream regions of the orthologous genes were aligned it seemed that some of the genes are preceded by a different but clearly distinguishable box (not shown). The protein encoded by *ykbA* in *B. subtilis* has recently been shown to be a Ser/Thr exchanger and was consequently renamed SteT (37). A similar functionality of the protein ortholog in the members of the *Bacillus cereus* group is supported by the codon identification of the T-box.

The NhaC family

In contrast to the APC family transporters the NhaC family is rather small and Firmicutes contain only one (*e.g. Clostridium acetobutylicum*, *Staphylococci*), two (*e.g. C. difficile*, *E. faecalis,* several *Bacilli* and *Lactobacilli*) or three (e.g. *B. cereus*-group) homologs (38). Moreover, in most of these species the expression of one of the homologous genes is controlled by a Tyr T-box and in *L. plantarum* both paralogous genes are preceded by such a T-box. In contrast, the homologous genes in *B. subtilis*, *B. licheniformis* and *B. clausii* lack a T-box and the same holds for the homologous genes in the Staphylococci. However, the latter genes do seem preceded by a different but clearly distinguishable box (not shown). The presence of a Tyr T-box suggests the related proteins play a role in the import, or eventually the production, of tyrosine. Surprisingly, the NhaC family homologs present in *B. firmus* (gene *nhaC*) (39,40) and *B. subtilis* (genes *yheL* and *yqkI*) (41) have been identified as Na+/H+ antiporter (NhaC/YheL) or malic acid/sodium lactate antiporter (YqkI), respectively. It was not directly clear how the latter functionalities should be connected to the uptake or synthesis of tyrosine. The T-box regulated NhaC family homologs are most similar to the *B. firmus* NhaC and thus are expected to display similar/identical molecular function (i.e. Na+/H+ antiport). However, considering the fact that the malic acid/sodium lactate antiporter of the same family has Na+/H+ antiport functionality (41) and that NhaC of *B. firmus* was not extensively characterized for substrate specificity (39,40) it could very well be that the latter transporter operates as an acid/ salt antiporter, where the acid is very similar to malic acid in structure and the salt to sodium lactate. Given the fact that one of the final steps in the biosynthesis of tyrosine is a transamination reaction with glutamate or aspartate as source of the amino-group and that aspartic acid is structurally very similar to malic acid, this hypothesis is appealing. While such a functional annotation remains speculative, the presence of the T box does imply a biological role in amino acid biosynthesis instead of Na+/H+ homeostasis as was previously suggested (41).

The LIVCS family

The *Bacilli* of the *Bacillus cereus* group, the *Lactobacilli* and the *Clostridia* contain several branched-chain amino acid cation symporters of the LIVCS-family (38), some of which are

T-box regulated. In *L. plantarum*, for instance, two out of three are T-box regulated with specifier codons corresponding to valine and isoleucine. Since the three branched-chain amino acids share very similar molecular properties (e.g. size and hydrophobicity) we expect that these transporters are not highly specific despite their proposed amino acid specific control, but merely that expression of the "multi-specific" system has been brought under the control of the individual amino acids. Indeed, the orthologous transporters that have been characterized in *L. delbrueckii* (BrnQ; (35)), *C. glutamicum* (BrnQ; (42,43)) and *P. aeruginosa* (BraZ; (44) displayed transport of all three branched-chain amino acids. Surprisingly, the species of the *Bacillus cereus* group carry 6 LIVCS homologs, which seems to be rather redundant as there are only three branched-chain amino acids. However, one of the homologs appears to be regulated by a Thr T-box and thus seems to have specialized in a different direction. Threonine is very similar in size to isoleucine but is hydrophilic instead of hydrophobic. One could imagine that only a few amino acid replacements could be sufficient to accommodate the transport of molecules of different polarity (i.e. adding a few charged/hydrophilic residues).

The NSS family

Finally, one of the experimentally characterized branched-chain amino acid transporters of the LIVCS family, BraZ of *P. aeruginosa*, was shown to have a clear preference for isoleucine and valine over leucine (44). In this respect it is noteworthy that in the *Bacilli* of the *Bacillus cereus* group the expression of one of the homologs of the NSS family (neurotransmitter:sodium symport) is controlled by a Leu T-box, whereas such a box is lacking for the LIVCS homologs in those species. Besides the Leu T-box regulated NSS transporter, the *Bacilli* of the *Bacillus cereus* group contain three other homologs of the same family, two of which are controlled by a Trp and a Phe T-box, respectively. These two amino acids agree well with the experimentally determined functionality of the NSS homolog TnaT in *Symbiobacterium thermophilum* (45).

Hypothetical proteins controlled by a T-box.

Another class of proteins to which significant functional information could be added (when compared to the NCBI-annotation) using the specificity of the detected T-box is that of the so-called hypothetical proteins (data accumulated in Table 3). Obviously, when orthologous proteins in related species were also of unknown function, specifier codon information clearly improved their annotation. We compared the T-box predicted amino acid specificity to the functional prediction present in the ERGO database (46) and found that for 8 out of 19 proteins the ERGO database contained a specific annotation which complied with our information. In the other 11 cases the proteins were annotated as hypothetical protein in the ERGO database.

On the origin and evolution of the various T-boxes

Two evolutionary interesting findings came out of the analysis described in the previous section. Both findings, which will be discussed below, implicated every T-box element as a connected yet independent "functional module", in the sense that it co-evolved with the gene/operon it regulates, but at the same time could have been duplicated independently of that gene/operon and could be acquired to regulate another gene/operon. To study the putative evolution of T-boxes further, the T-boxes preceding the genes that encode the t-RNA-ligases for the branched-chain amino acids (*ileS, leuS and valS*) were collected. These were chosen because T-box regulation of these genes is most wide-spread (see Figure1) and because the genes themselves

Table 3: Proposed annotation of the genes regulated by a T-box that were assigned as "hypothetical protein" in the original NCBI annotation file.

Species	Gene ID	T-box	proposed function
Bacillus halodurans C-125	*BH0807*	Lys	Lysine-specific permease
Bacillus subtilis 168	*BSU02530*	Trp	Anti TRAP protein
	BSU34010 (yvbW)	Leu	Leucine-specific permease
Enterococcus faecalis V583	*EF2480*	Gly	Gly related hypothetical
Lactobacillus acidophilus NCFM	*LBA1071*	Ile	Ile related hypothetical
Lactobacillus johnsonii NCC 533	*LJ0632*	Met	5-methyltetrahydropteroyltriglutamate—homocysteine methyltransferase (Methionine synthase)
Lactobacillus plantarum WCFS1	*lp_3283*	Met	5-methyltetrahydropteroyltriglutamate—homocysteine methyltransferase (Methionine synthase)
Listeria sp. [1]	*lmo1740*	His	Histidine transport system permease protein hisM
	lmo2587	Met	Met related cytosolic hypothetical
Staphylococcus aureus [2]	*SA0347*	Met	Cystathionine gamma-synthase
	SA1199	Trp	Anthranilate synthase component I
Streptococcus agalactiae [3]	*SAG0809*	Ala	Ala-tRNA ligase related hypothetical
Streptococcus pneumoniae [4]	*spr0489*	Val	Val-tRNA ligase related hypothetical
	spr1241	Ala	Ala-tRNA ligase related hypothetical
	spr1331	Gly	Gly-tRNA ligase related hypothetical
	spr1471	Thr	Thr-tRNA ligase related hypothetical
	spr1638	Trp	Trp biosynthesis related hypothetical
Streptococcus thermophilus [5]	*str0474*	Val	Val-tRNA ligase related hypothetical
	str1594	Trp	Chorismate mutase

1 Listeria innocua Clip11262*, Listeria monocytogenes* EGD-e*, L. monocytogenes* 4bF2365*. 2 Staphylococcus aureus* MW2*, S. aureus* N315*. 3 Streptococcus agalactiae* 2603*, S. agalactiae* A909*, S. agalactiae* NEM316*. 4 Streptococcus pneumoniae* TIGR4*, S. pneumoniae* R6*. 5 Streptococcus thermophilus* CNRZ106*, S. thermophilus* LMG18311*.*

are phylogenetically most closely related (Figure 4A). The corresponding T-boxes were aligned and a bootstrapped Neighbor-Joining tree was created to visualize the similarity between the various T-boxes (Figure 4B). In first instance, we expected Figures 4A and 4B to be similar because of the implicit "conservation" of T-box regulation between the species. In contrast, we observed that it was hard to distinguish between the various amino acid specific T-boxes in the NJ-tree of the T-boxes, *e.g.* in several cases different amino acid specificities clustered together (see Figure 4B). Smaller homogeneous clusters were obtained but only for closely related species (for instance for Leu and Val in the *bacilli*). This finding indicated that, although co-evolution with the regulated gene should occur, other processes should have been at play to blur the picture. In addition, the results were in line with our inability earlier to derive clear amino acid specific overall T-box motifs and confirmed the idea that any additional amino acid specificity of the T-box sequence apart from the specifier-codon would be hard to find (when it would be there at all). Moreover, it pointed out that the NJ-tree for the T-box elements should not be interpreted as a representation of the evolution per se. On the other hand, it also testified that when the sequences do cluster in a reliable cluster in such a tree (high bootstrap values) that the chance that they are evolutionary related is very high.

To limit possible obscuring effects of comparing sequences between species, we collected and compared the T-box sequences within one species. *B. anthracis* was chosen as an example because its genome had a relatively high amount of T-boxes that could be compared. Furthermore, for clarity the analysis was restricted to the T-boxes that accompany transport systems and the related tRNA ligases. We mentioned earlier that the association of transport systems with T-boxes is characterized by a restricted occurrence and high species-specificity. An excellent example of that is provided by members of the APC and MFS transporter families. Although these families are very large, there was only one family-member found regulated by a T-box and only in bacilli. Moreover, it appeared in the case of the APC family that the particular family member that was regulated depended on the lineage within the bacilli (*B. subtilis* and *B. licheniformis* versus *B. cereus* group) and that within the respective orthologous group of genes only a limited number of the genes were so regulated (Table 2). The only likely scenario to explain the phylogenetically limited occurrence of the transporter T-box associations is that of a late acquisition of the regulatory element. All other scenarios would imply massive loss of the T-box regulation within all transporter families. The former view is supported by the analysis in *B. anthracis* given in Figure 4C and the analysis of transporter associated T-boxes in the lactobacilli *L. plantarum* and *L. acidophilus* (not shown). For instance, the depicted multiple sequence alignment and NJ-tree (Figure 4C) for *B. anthracis* are highly suggestive of a direct relationship between the Thr T-box found in front of one of the LIVCS systems and the one found in front of the Thr-tRNA ligase. Likewise, a similar relationship should exist between the Ile T-box found in front of another LIVCS homolog and that of the Ile-tRNA ligase. In contrast, a direct evolutionary relationship between the two LIVCS associated T-boxes could be ruled-out on basis of the same observations.

The results presented in Figure 4 are suggestive of yet another way in which the T-boxes should have evolved. Both the NJ-tree and the multiple sequence alignment relate the Phe T-box found in front of one of the NSS family transporters to the Tyr T-box associated with the Tyr-tRNA ligase. In fact, the similarity between these boxes is even higher than between the Tyr T-boxes preceding the gene encoding an NhaC family transporter and the Tyr-tRNA ligase (Figure 4C). It thus seems that the Tyr T-box of the tRNA ligase was duplicated -as this T-box is present in various Firmicutes species- and has diverged/adapted to control a Phe transporter in the *Bacilli* of the *B. cereus* group. In fact, it has been shown experimentally that a single nucleotide change in the specifier codon of the Tyr T-box of *tyrS* in *B. subtilis* was enough to change the amino acid specificity to phenylalanine (8). In addition, a follow-up study showed that 9 other changes of specificity could be achieved by altering the specifier codon of *tyrS,* although in some cases the mutation in the specifier codon was combined with a nucleotide change in the antiterminator sequence (9). Similarly, it can be assumed that the Tyr T-box preceding the gene encoding the NhaC family transporter to have derived from the Tyr-tRNA ligase T-box. However, as this event seems to have occurred earlier -the Tyr T-box control of the NhaC ortholog is present in several Firmicutes- the sequences had more time to diverge.

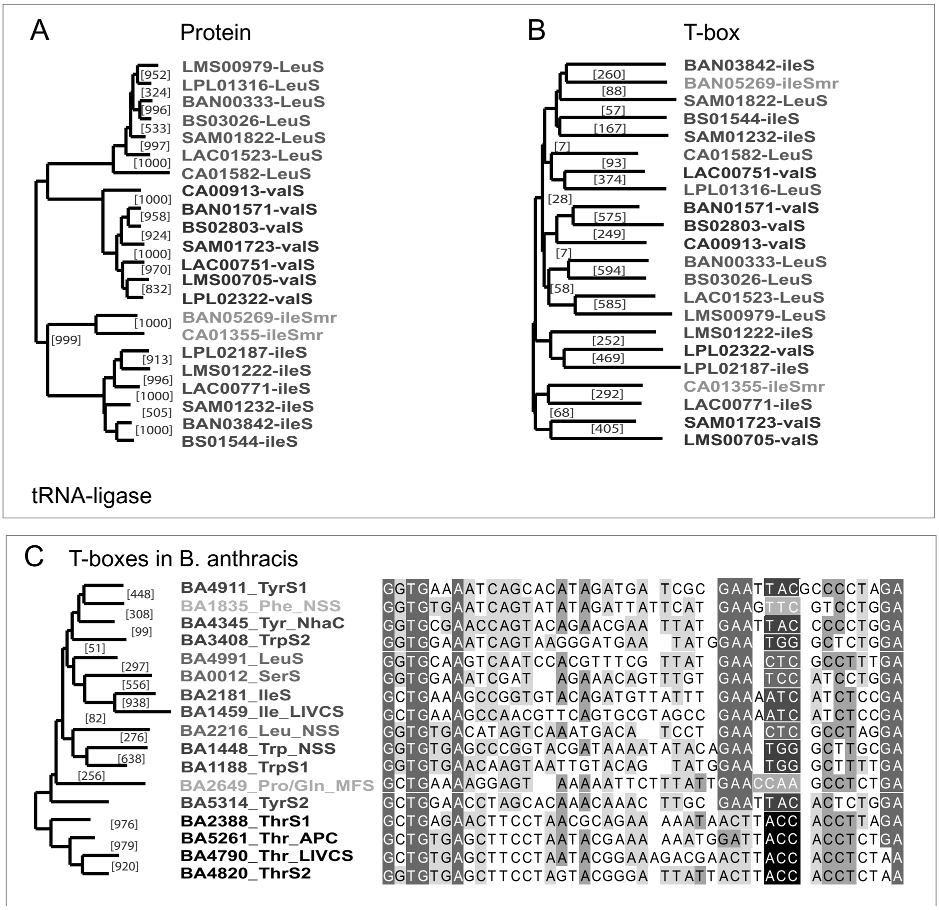

A
Protein
tRNA-ligase
[952]
[324]
[996]
[533]
[997]
[1000]
[1000]
[958]
[924]
[1000]
[970]
[832]
[1000]
[999]
[913]
[996]
[1000]
[505]
[1000]
LMS00979-LeuS
LPL01316-LeuS
BAN00333-LeuS
BS03026-LeuS
SAM01822-LeuS
LAC01523-LeuS
CA01582-LeuS
CA00913-valS
BAN01571-valS
BS02803-valS
SAM01723-valS
LAC00751-valS
LMS00705-valS
LPL02322-valS
BAN05269-ileSmr
CA01355-ileSmr
LPL02187-ileS
LMS01222-ileS
LAC00771-ileS
SAM01232-ileS
BAN03842-ileS
BS01544-ileS
B
T-box
[260]
[88]
[57]
[167]
[7]
[93]
[374]
[28]
[575]
[249]
[7]
[594]
[58]
[585]
[252]
[469]
[292]
[68]
[405]
BAN03842-ileS
BAN05269-ileSmr
SAM01822-LeuS
BS01544-ileS
SAM01232-ileS
CA01582-LeuS
LAC00751-valS
LPL01316-LeuS
BAN01571-valS
BS02803-valS
CA00913-valS
BAN00333-LeuS
BS03026-LeuS
LAC01523-LeuS
LMS00979-LeuS
LMS01222-ileS
LPL02322-valS
LPL02187-ileS
CA01355-ileSmr
LAC00771-ileS
SAM01723-valS
LMS00705-valS
C
T-boxes in B. anthracis
[448]
[308]
[99]
[51]
[297]
[556]
[938]
[82]
[276]
[638]
[256]
[976]
[979]
[920]
BA4911_TyrS1 GGTGAAAATCAGCACATAGATGA TCGC GAATTACGCCCTAGA
BA1835_Phe_NSS GGTGTGAATCAGTATATAGATTATTCAT GAAGTTC GTCCTGGA
BA4345_Tyr_NhaC GGTGCGAACCAGTACAGAACGAA TTAT GAATTAC CCCCTGGA
BA3408_TrpS2 GGTGGAAATCAGTAAGGGATGAA TATGGAA TGG GCTCTGGA
BA4991_LeuS GGTGCAAGTCAATCCACGTTTCG TTAT GAA CTC GCCTTTGA
BA0012_SerS GGTGGAAATCGAT AGAAACAGTTTGT GAA TCC ATCCTGGA
BA2181_IleS GCTGAAAGCCGGTGTACAGATGTTATTT GAAAATC ATCTCCGA
BA1459_Ile_LIVCS GCTGAAAGCCAACGTTCAGTGCGTAGCC GAAAATC ATCTCCGA
BA2216_Leu_NSS GGTGTGACATAGTCAAATGACA TCCT GAA CTC GCCTAGGA
BA1448_Trp_NSS GGTGTGAGCCCGGTACGATAAAATATACAGAA TGG GCTTGCGA
BA1188_TrpS1 GGTGTGAACAAGTAATTGTACAG TATGGAA TGG GCTTTTGA
BA2649_Pro/Gln_MFS GCTGAAAAGGAGT AAAAATTCTTTATTGAACCAA GCCTCTGA
BA5314_TyrS2 GCTGGAACCTAGCACAAACAAAC TTGC GAATTAC ACTCTGGA
BA2388_ThrS1 GCTGAGAACTTCCTAACGCAGAA AAATAACTTACC ACCTTAGA
BA5261_Thr_APC GCTGTGAGCTTCCTAATACGAAA AAATGGATTACC ACCTCTGA
BA4790_Thr_LIVCS GCTGTGAGCTTCCTAATACGGAAAGACGAACTTACC ACCTCTAA
BA4820_ThrS2 GGTGTGAGCTTCCTAGTACGGGA TTATTACTTACC ACCTCTAA

Figure 4 (a color version of the figure can be found at page 176): The evolutionary relationship between some T-boxes. (A) shows a putative phylogeny of the branched-chain amino acid tRNA ligases of *B. anthracis* Ames, *B. subtilis* 168, *C. acetobutylicum* ATCC824D, *L. acidophilus* NCFM, *L. plantarum* WCFS1, *L. mesenteroides* ATCC8293 and *S. aureus* Mu50. (B) shows the Neighbor Joining tree for the related T-boxes. This tree does not reflect the true phylogeny of the regulatory elements but merely serves as an indicator of element similarity. (C) shows the Neighbor Joining tree for various T-boxes found in *B. anthracis* Ames next to the specifier codon containing part of the corresponding multiple sequence alignment. The specifier codon is indicated in white letters. The amino acid specificity is color-coded: red and orange relate to Ile, green to Leu, light blue to Phe, beige to Pro or Gln, pink to Ser, brown to Thr, turquoise to Trp, purple to Tyr and dark blue to Val. The functional group of the regulated gene is indicated by the letters that follow the amino acid code. The group can be either a transporter of the APC, LIVCS, MFS, NhaC or NSS family or a tRNA-ligase (S or Smr for mupirocin-resistant tRNA ligase).The NSS-family transport proteins regulated by a Leu, Phe and Trp T-box are in-paralogs characteristic for the species of the *Bacillus cereus* group. The purple numbers between brackets indicate the bootstrap support for the displayed clusters in the Neighbor Joining trees (out of 1000).

Discussion and Conclusion

The sequence signature of a T-box is very specific and as a result T-boxes can be readily identified. Using specific T-box HMMs and RNA-structure predictions, we identified a large number of T-boxes and their amino acid specificity in sequenced prokaryote genomes. Inspection of the upstream regions of the genes encoding the tRNA ligases and several orthologous transporters in the *Firmicutes* suggests that our prediction is essentially complete; no sequential or structural clues of T-box(-like) origin were observed in any of the upstream regions of orthologous genes that were not predicted to be T-box regulated. Although a substantial number of the T-boxes had been described before, many others are described here for the first time, like most of the ones related to amino acid transport. Clear phylogenetic distribution patterns for T-box regulation within the different bacterial lineages are observed. For instance, in *Streptococci* only the expression of tRNA ligases is controlled by T-boxes. In contrast, in *Lactobacilli*, many genes involved in tRNA ligation, transport or biosynthesis of amino acids seem to be controlled using T-boxes. One of the most telling examples in this respect is related to methionine. Our analysis suggests that in the absence of methionine, *L. plantarum* uses a single mechanism to switch on not only a transport system for the amino acid itself, but also for the precursors and co-factors needed in its biosynthesis (see additional file 1). Among *Bacillus* species, the *cereus* group appears to be distinct by their tendency to regulate various amino acid biosynthesis pathways and amino acid transport systems using T-boxes, whereas other *Bacilli*, like *B. subtilis,* seem to employ other, sometimes more complex, regulatory schemes to control the same pathways and systems.

We could use the prediction of the amino acid specificity of various T-boxes to improve the functional annotation of a large number of genes. In particular the functional annotation of genes related to amino acid transport and genes with unknown substrate specificity, genes for which it is normally quite difficult to find functional attributes, could be improved considerably. In our opinion, the procedure of improving annotation through knowledge of the regulatory signals can be generalized and should be used on a much broader scale than currently is being done.

Most amino acids are encoded by multiple codons. Leu for instance, is encoded by six different codons (CUA, CUU, CUG, CUC, UUA and UUG). Remarkably, T-boxes seem to have a conserved preference for certain codons within as well as between species (supplementary Table S1). Evaluation of these preferences shows that they comply almost perfectly with the rules observed by Elf et al. for the codon usage by *E. coli* (47). In an elegant study these authors analyzed the dependence of the charging of various codon-specific tRNAs on the use of various codons in particular proteins. They concluded that: “when codon reading is part of a control loop that regulates synthesis of missing amino acid, the translation rate of the selected codon should be as sensitive as possible to starvation” (47). Indeed we find that for all but one of the most predominantly used codons in T-boxes, the sensitivity for depletion observed by (47) was among the highest. The only exception was the T-box codon for Ala (GCU). Therefore, in case the conclusions by Elf et al. are also valid for Gram-positive bacteria, our findings suggest that the codons that are sensitive to depletion are preferentially used in T-box regulation.

Although no clear phylogenetic distribution was observed for T-box regulation per se, some trends are apparent. In general, ecologically flexible organisms (e.g. *C. acetobutylicum, L. plantarum, Bacilli of the B. cereus group*) within each family of species have more T-box regulated genes in comparison to organisms with only one or a few specific natural habitats (e.g. *L. johnsonii* and *B. licheniformis*). In addition, pathogenic species with a strictly determined niche are generally found to have far less T-box regulated genes (e.g. *Mollicutes*, *S. pneumoniae*). These differences were observed in all classes of T-box regulated genes (tRNA ligation, biosynthesis and transport).

Riboswitches are considered to be among the oldest regulatory systems in bacteria because of their independence of regulatory proteins and widespread biological distribution (14). One may therefore expect that T-boxes are abundantly present among all different lineages of bacteria which is not the case based on the results from our and other studies (10,29). In fact, these regulatory elements can only be found in a few bacterial phyla and only abundantly in the phylum *Firmicutes*. This implies that either *Firmicutes* developed T-box regulation after their branching off from the other bacteria or that the other bacteria lost the system soon after the branching off of the *Firmicutes* to evolve more complex regulatory systems. Which of the two scenarios is most likely remains unclear. The fact that T-box regulation is RNA-based argues in favor of the latter scenario, whereas the abundance, versatility and elegance of T-box regulation makes it hard to envisage all other bacteria should have lost the system. The advantage of a more complex system is that it is capable of processing multiple input signals. The development of more complex systems is also observed in *Firmicutes*. An illustrative example is provided by the regulation of Trp biosynthesis in *B. subtilis*. It was shown before that in this species tryptophan biosynthesis is repressed by the TRAP protein (MtrB), which is inactivated upon binding of tryptophan (for a review, see (48)). The TRAP protein activity is repressed via the anti-TRAP protein (BSU02530), which is in its turn preceded by a tryptophan T-box (11,49). Therefore, TRAP regulates the expression of Trp biosynthesis genes by combining both the Trp-tRNA depletion level as well as Trp concentration to one regulatory output signal (11). TRAP is found only in a phylogenetically closely related subset of *Firmicutes* that are all part of the family of *Bacillaceae* (28). It is therefore likely that the T-box in *Bacillaceae* was replaced by the more complex TRAP system.

Although the origin of T-box elements in prokaryotes remains unclear, our results do provide a view on how they could have developed. We hypothesize that the T-boxes may have evolved in four clearly distinct ways as depicted in Figure 5: ia) co-evolution with the regulated gene or operon; ib) co-evolution and divergence with the regulated gene or operon to adopt a new specificity; iia) duplication and insertion of the regulatory element in front of a gene or operon that encodes functions related to the T-box-specified amino acid. iib) duplication and divergence toward a new amino acid specificity after duplication. Indeed, we have found (strong) indications for all four processes. Overall T-box regulation is most wide-spread for the tRNA ligases and especially in case of the branched chain amino acids Leu, Ile and Val. As these genes are also phylogenetically closely related we presume the ancestral T-box to be related to one of these amino acids and related to one of these t-RNA-ligases. After the conception of the T-box elements, they appear to have spread according to the scenarios sketched above.

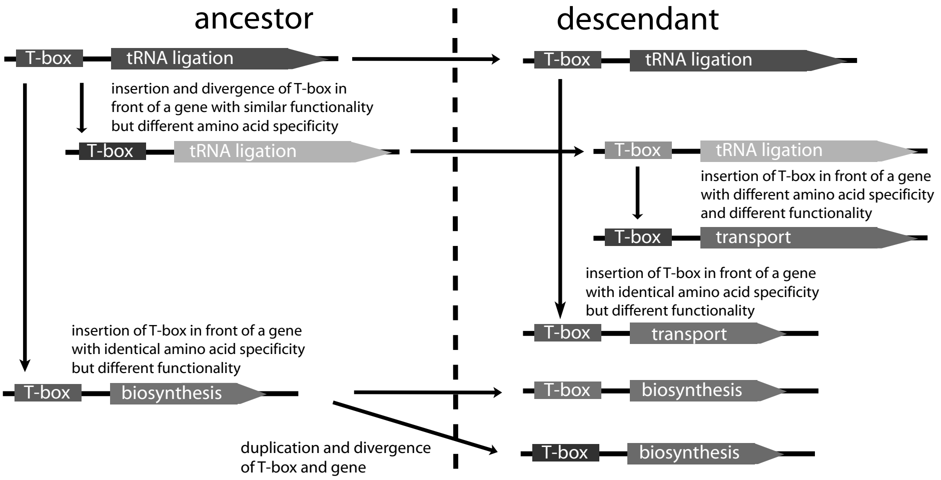

ancestor
descendant
T-box
tRNA ligation
insertion and divergence of T-box in front of a gene with similar functionality but different amino acid specificity
insertion of T-box in front of a gene with different amino acid specificity and different functionality
transport
insertion of T-box in front of a gene with identical amino acid specificity but different functionality
insertion of T-box in front of a gene with identical amino acid specificity but different functionality
biosynthesis
duplication and divergence of T-box and gene

Figure 5 (a color version of the figure can be found at page 177): Scenarios for T-box evolution.
Most T-boxes are linked on the genome to genes that encode proteins related to three main functional categories, namely: tRNA-ligation, amino acid transport and amino acid biosynthesis. Our results suggest that there are several ways in which these T-boxes (amino acid specificity is color-coded) evolve: T-boxes are duplicated and inserted in front of genes from the same functional category but with different amino acid specificity and, vice versa, in front of genes with the same amino acid specificity but from another functional category. In addition, duplication and insertion in front of genes from another functional category with different amino acid specificity takes place. These different routes impose convergent as well as divergent constraints on the evolution of the T-box sequence. As shown, already within a few speciations a plethora of T-box sequences could develop from one initial sequence. In effect, these processes assure a more or less independent evolution of the T-boxes from the genes they control.

References

1. **Winkler, W.C. (2005)** Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr Opin Chem Biol, 9, 594-602.*
2. **Deutscher, J., Francke, C. and Postma, P.W. (2006)** How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiol Mol Biol Rev, 70, 939-1031.*
3. **Schilling, O., Langbein, I., Muller, M., Schmalisch, M.H. and Stulke, J. (2004)** A protein-dependent riboswitch controlling ptsGHI operon expression in *Bacillus subtilis*: RNA structure rather than sequence provides interaction specificity. *Nucleic Acids Res, 32, 2853-2864.*
4. **Grundy, F.J., Yousef, M.R. and Henkin, T.M. (2005)** Monitoring uncharged tRNA during transcription of the *Bacillus subtilis glyQS* gene. *J Mol Biol, 346, 73-81.*
5. **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.*
6. **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.*
7. **Grundy, F.J. and Henkin, T.M. (2003)** The T box and S box transcription termination control systems. *Front Biosci, 8, d20-31.*
8. **Grundy, F.J. and Henkin, T.M. (1993)** tRNA as a positive regulator of transcription antitermination in *B. subtilis. Cell, 74, 475-482.*
9. **Grundy, F.J., Hodil, S.E., Rollins, S.M. and Henkin, T.M. (1997)** Specificity of tRNA-mRNA interactions in *Bacillus subtilis* tyrS antitermination. *J Bacteriol, 179, 2587-2594.*
10. **Seliverstov, A.V., Putzer, H., Gelfand, M.S. and Lyubetsky, V.A. (2005)** Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiol, 5, 54.*
11. **Sarsero, J.P., Merino, E. and Yanofsky, C. (2000)** A *Bacillus subtilis* operon containing genes of unknown function senses tRNATrp charging and regulates expression of the genes of tryptophan biosynthesis. *Proc Natl Acad Sci U S A, 97, 2656-2661.*

12. **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.*

13. **Panina, E.M., Vitreschak, A.G., Mironov, A.A. and Gelfand, M.S. (2003)** Regulation of biosynthesis and transport of aromatic amino acids in low-GC Gram-positive bacteria. *FEMS Microbiol Lett, 222, 211-220.*

14. **Vitreschak, A.G., Rodionov, D.A., Mironov, A.A. and Gelfand, M.S. (2004)** Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet, 20, 44-50.*

15. **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.*

16. **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.*

17. **Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004)** WebLogo: a sequence logo generator. *Genome Res, 14, 1188-1190.*

18. **Edgar, R.C. (2004)** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32, 1792-1797.*

19. **Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997)** The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res,* **25**, 4876-4882.

20. **Durbin R, E.S., Krogh A, Mitchison G. (1998)** *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge.

21. **Zuker, M. and Jacobson, A.B. (1998)** Using reliability information to annotate RNA secondary structures. *Rna, 4, 669-679.*

22. **Grundy, F.J. and Henkin, T.M. (1994)** Conservation of a transcription antitermination mechanism in aminoacyl-tRNA synthetase and amino acid biosynthesis genes in gram-positive bacteria. *J Mol Biol, 235, 798-804.*

23. **Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005)** Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res, 33, D121-124.*

24. **Henkin, T.M., Glass, B.L. and Grundy, F.J. (1992)** Analysis of the *Bacillus subtilis tyrS* gene: conservation of a regulatory sequence in multiple tRNA synthetase genes. *J Bacteriol, 174, 1299-1306.*

25. **Vasconcelos, A.T., Ferreira, H.B., Bizarro, C.V., Bonatto, S.L., Carvalho, M.O., Pinto, P.M., Almeida, D.F., Almeida, L.G., Almeida, R., Alves-Filho, L. et al. (2005)** Swine and poultry pathogens: the complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*. *J Bacteriol, 187, 5568-5577.*

26. U**eda, K., Yamashita, A., Ishikawa, J., Shimada, M., Watsuji, T.O., Morimura, K., Ikeda, H., Hattori, M. and Beppu, T. (2004)** Genome sequence of *Symbiobacterium thermophilum,* an uncultivable bacterium that depends on microbial commensalism. *Nucleic Acids Res, 32, 4937-4944.*

27. **Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S., Lapidus, A., Goltsman, E., Mazur, M., Pusch, G.D. et al. (2004)** Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. *Nat Biotechnol, 22, 1554-1558.*

28. **Gutierrez-Preciado, A., Jensen, R.A., Yanofsky, C. and Merino, E. (2005)** New insights into regulation of the tryptophan biosynthetic operon in Gram-positive bacteria. *Trends Genet, 21, 432-436.*

29. **Henkin, T.M. (1996)** Control of transcription termination in prokaryotes. *Annu Rev Genet, 30, 35-57.*

30. **Tojo, S., Satomura, T., Morisaki, K., Deutscher, J., Hirooka, K. and Fujita, Y. (2005)** Elaborate transcription regulation of the *Bacillus subtilis ilv-leu* operon involved in the biosynthesis of branched-chain amino acids through global regulators of CcpA, CodY and TnrA. *Mol Microbiol, 56, 1560-1573.*

31. **Shivers, R.P. and Sonenshein, A.L. (2005)** *Bacillus subtilis ilvB* operon: an intersection of global regulons. *Mol Microbiol, 56, 1549-1559.*

32. **Grandoni, J.A., Fulmer, S.B., Brizzio, V., Zahler, S.A. and Calvo, J.M. (1993)** Regions of the *Bacillus subtilis ilv-leu* operon involved in regulation by leucine. *J Bacteriol,* **175**, 7581-7593.

33. **Saier, M.H., Jr. (2000)** A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol Mol Biol Rev, 64, 354-411.*

34. **Bandell, M., Ansanay, V., Rachidi, N., Dequin, S. and Lolkema, J.S. (1997)** Membrane potential-generating malate (MleP) and citrate (CitP) transporters of lactic acid bacteria are homologous proteins.

Substrate specificity of the 2-hydroxycarboxylate transporter family. *J Biol Chem, 272, 18140-18146.*

35. **Stucky, K., Hagting, A., Klein, J.R., Matern, H., Henrich, B., Konings, W.N. and Plapp, R. (1995)** Cloning and characterization of *brnQ*, a gene encoding a low-affinity, branched-chain amino acid carrier in *Lactobacillus delbrueckii subsp. lactis* DSM7290. *Mol Gen Genet, 249, 682-690.*
36. **Jack, D.L., Paulsen, I.T. and Saier, M.H. (2000)** The amino acid/polyamine/organocation (APC) superfamily of transporters specific for amino acids, polyamines and organocations. *Microbiology, 146 (Pt 8), 1797-1814.*
37. **Reig, N., del Rio, C., Casagrande, F., Ratera, M., Gelpi, J.L., Torrents, D., Henderson, P.J., Xie, H., Baldwin, S.A., Zorzano, A. et al. (2007)** Functional and structural characterization of the first prokaryotic member of the L-amino acid transporter (LAT) family: a model for APC transporters. *J Biol Chem, 282, 13270-13281.*
38. **Ren, Q., Kang, K.H. and Paulsen, I.T. (2004)** TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res, 32, D284-288.*
39. **Ivey, D.M., Guffanti, A.A., Bossewitch, J.S., Padan, E. and Krulwich, T.A. (1991)** Molecular cloning and sequencing of a gene from alkaliphilic *Bacillus firmus* OF4 that functionally complements an *Escherichia coli* strain carrying a deletion in the *nhaA* Na+/H+ antiporter gene. *J Biol Chem, 266, 23483-23489.*
40. **Ito, M., Guffanti, A.A., Zemsky, J., Ivey, D.M. and Krulwich, T.A. (1997)** Role of the *nhaC*-encoded Na+/H+ antiporter of alkaliphilic *Bacillus firmus* OF4. *J Bacteriol, 179, 3851-3857.*
41. **Wei, Y., Guffanti, A.A., Ito, M. and Krulwich, T.A. (2000)** *Bacillus subtilis* YqkI is a novel malic/Na+-lactate antiporter that enhances growth on malate at low protonmotive force. *J Biol Chem, 275, 30287-30292.*
42. **Ebbighausen, H., Weil, B. and Kramer, R. (1989)** Transport of branched-chain amino acids in *Corynebacterium glutamicum*. *Arch Microbiol, 151, 238-244.*
43. **Tauch, A., Hermann, T., Burkovski, A., Kramer, R., Puhler, A. and Kalinowski, J. (1998)** Isoleucine uptake in *Corynebacterium glutamicum* ATCC 13032 is directed by the brnQ gene product. *Arch Microbiol, 169, 303-312.*
44. **Hoshino, T., Kose-Terai, K. and Uratani, Y. (1991)** Isolation of the braZ gene encoding the carrier for a novel branched-chain amino acid transport system in *Pseudomonas aeruginosa* PAO. *J Bacteriol, 173, 1855-1861.*
45. **Androutsellis-Theotokis, A., Goldberg, N.R., Ueda, K., Beppu, T., Beckman, M.L., Das, S., Javitch, J.A. and Rudnick, G. (2003)** Characterization of a functional bacterial homologue of sodium-dependent neurotransmitter transporters. *J Biol Chem, 278, 12703-12709.*
46. **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.*
47. **Elf, J., Nilsson, D., Tenson, T. and Ehrenberg, M. (2003)** Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science, 300, 1718-1722.*
48. **Babitzke, P. and Gollnick, P. (2001)** Posttranscription initiation control of tryptophan metabolism in *Bacillus subtilis* by the trp RNA-binding attenuation protein (TRAP), anti-TRAP, and RNA structure. *J Bacteriol, 183, 5795-5802.*
49. **Henkin, T.M. (2000)** Transcription termination control in bacteria. *Curr Opin Microbiol, 3, 149-153.*
50. S**trauch, M.A., Zalkin, H. and Aronson, A.I. (1988)** Characterization of the glutamyl-tRNA(Gln)-to-glutaminyl-tRNA(Gln) amidotransferase reaction of Bacillus subtilis. *J Bacteriol, 170, 916-920.*
51. **Curnow, A.W., Ibba, M. and Soll, D. (1996)** tRNA-dependent asparagine formation. *Nature, 382, 589-590.*

CHAPTER 6

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

Michiel Wels
Christof Francke
Robert Kerkhoven
Michiel Kleerebezem
Roland J. Siezen.

**Published in Nucleic Acids Res.,
Apr 2006; 34: 1947 - 1958.**

Acknowledgements
Thanks to Douwe Molenaar, Mariela Serrano 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.

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 feed-back 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* behaviour, for example 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 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). *L. 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 lactic acid bacteria (15), including other *Lactobacillaceae* like *Lactobacillus johnsonii* (16).

In this work, an in-depth phylogenetic foot-printing 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 foot-printing 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

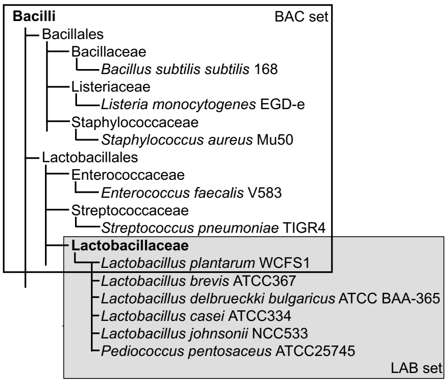

Bacilli
BAC set
Bacillales
Bacillaceae
Bacillus subtilis subtilis 168
Listeriaceae
Listeria monocytogenes EGD-e
Staphylococcaceae
Staphylococcus aureus Mu50
Lactobacillales
Enterococcaceae
Enterococcus faecalis V583
Streptococcaceae
Streptococcus pneumoniae TIGR4
Lactobacillaceae
Lactobacillus plantarum WCFS1
Lactobacillus brevis ATCC367
Lactobacillus delbrueckki bulgaricus ATCC BAA-365
Lactobacillus casei ATCC334
Lactobacillus johnsonii NCC533
Pediococcus pentosaceus ATCC25745
LAB set

Figure 1: Phylogenetic relation of species gathered from the TaxBrowser at NCBI. Relations are based on 16 S 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.

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 *Lactobacillacae* (including *Pediococcus pentosaceus,* which is also considered to be a member of the family of *Lactobacillacae*). By using these different species sets, we identified and distinguished between regulatory motifs conserved among *Bacilli* and/or *Lactobacillacae*. All motifs 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*, L. brevis* ATCC367*, L. delbreuckii* ssp. *bulgaricus* ATCC BAA-365*, L. casei* ATCC334 and *Pediococcus pentosaceus* ATCC25745. If not available in public databases (Genbank: http://www.ncbi.nlm.nih.gov/), the data was 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 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.

Transcriptional unit (TU) prediction:

Transcriptional unit predictions have been performed before for several of these genomes (18), but not all. Therefore, transcriptional unit (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 1) adjacent genes were positioned on the same coding strand, 2) adjacent genes had an intergenic region < 100bp, and 3) 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

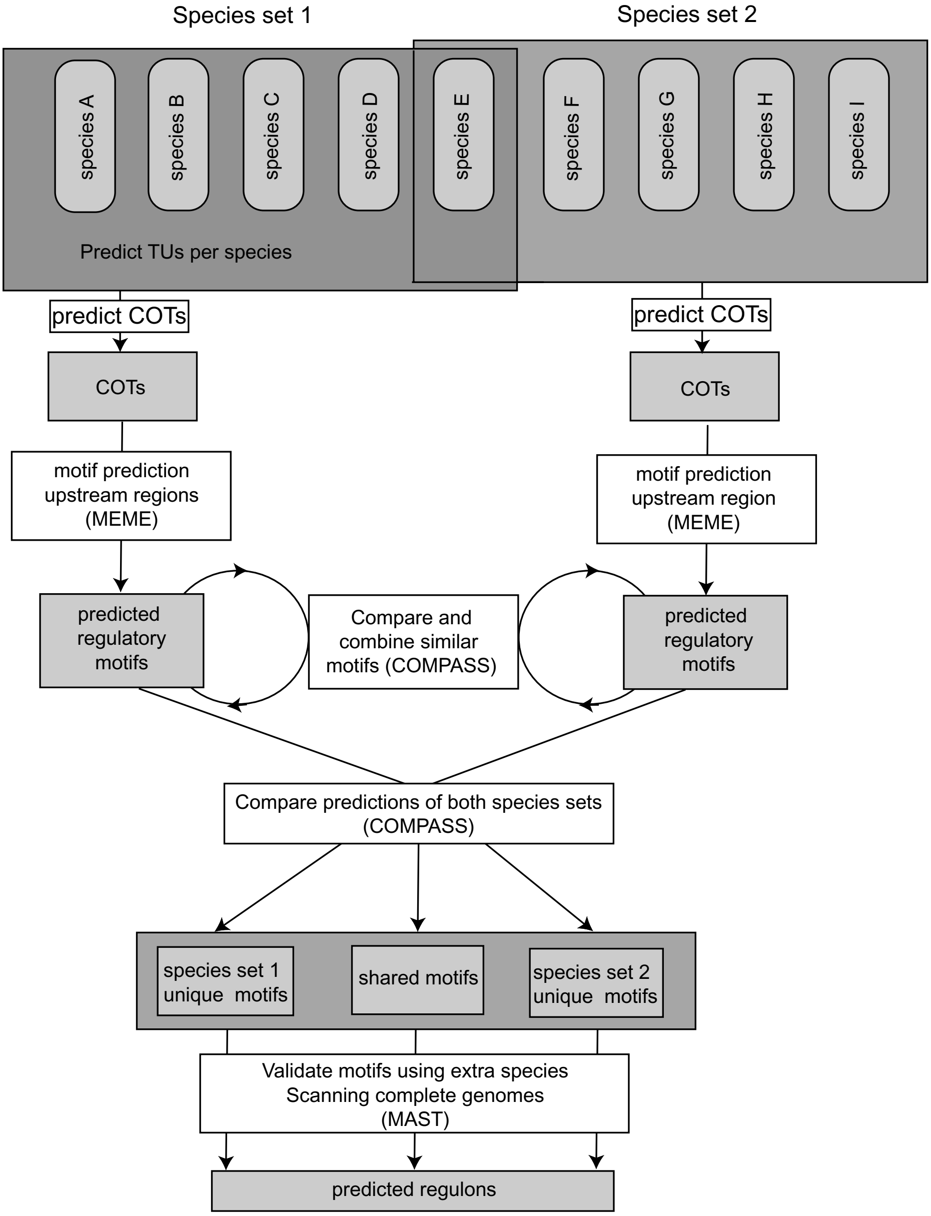

Species set 1
Species set 2
species A
species B
species C
species D
species E
species F
species G
species H
species I
Predict TUs per species
predict COTs
predict COTs
COTs
COTs
motif prediction upstream regions (MEME)
motif prediction upstream region (MEME)
predicted regulatory motifs
Compare and combine similar motifs (COMPASS)
predicted regulatory motifs
Compare predictions of both species sets (COMPASS)
species set 1 unique motifs
shared motifs
species set 2 unique motifs
Validate motifs using extra species Scanning complete genomes (MAST)
predicted regulons

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.

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 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 less than 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 bp of the coding sequence of the first gene of each transcriptional unit 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 twenty basepairs. 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: 1) MEME can predict motifs of different lengths, depending on a selected length range. If the length of a *cis*-acting

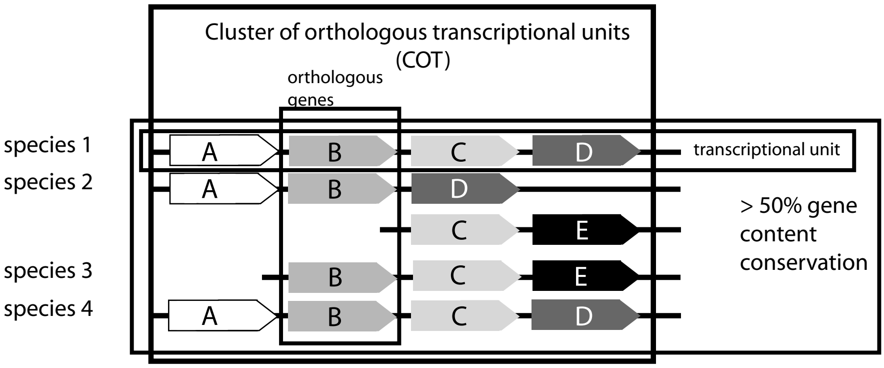


Figure 3: Prediction of the COTs (Clusters of orthologous TUs).
The TUs of the different species were compared. If >50% of the gene content of the smallest transcriptional unit was shared, the TUs were considered to be orthologous and combined into one cluster. Gene order was allowed to vary in the analysis.

element is not known, MEME will generate a motif with the correct size, and (2) 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 5 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 (due 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 preformed 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 below $10e^{-9}$ were considered positive. Hits with a p-value above $10e^{-9}$ but below $10e^{-5}$ 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 was 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 results). The

^{2}log(cy3/cy5) (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 to all correlations.

Results

Conserved orthologous transcriptional unit prediction

Transcriptional unit predictions were performed on all genomes of both species sets (Table 1). This led to the prediction of 9,618 TUs for the BAC set, with a mean of 1.86 genes per TU and 6,464 TUs, 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 approximately 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 to gene pairs that where predicted not to be present in the same TU. Figure 4 shows the distribution of expression correlations of predicted TU gene pairs compared to 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. 70% of all correlations within TUs are >0.50, while only 15% of the complete dataset correlates >0.50.

Table 1: Characteristics for each species set.

	BAC	LAB
Number of species	6	6
Number of genes	17922	11436
Genes/species (mean)	2987	1906
Number of TUs	9618	6464
Genes/transcriptional unit (mean)	1.86	1.77
TUs/species (mean)	1603	1077
Number of COTs	775	527
TUs/COT (mean) [1]	5.2	5.0
Number of unique motifs	195	195
(E-value < 0.1)		

[1] Since a transcriptional unit 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.

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 fewer 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 resulted in the identification of 390 different motifs, equally divided over both

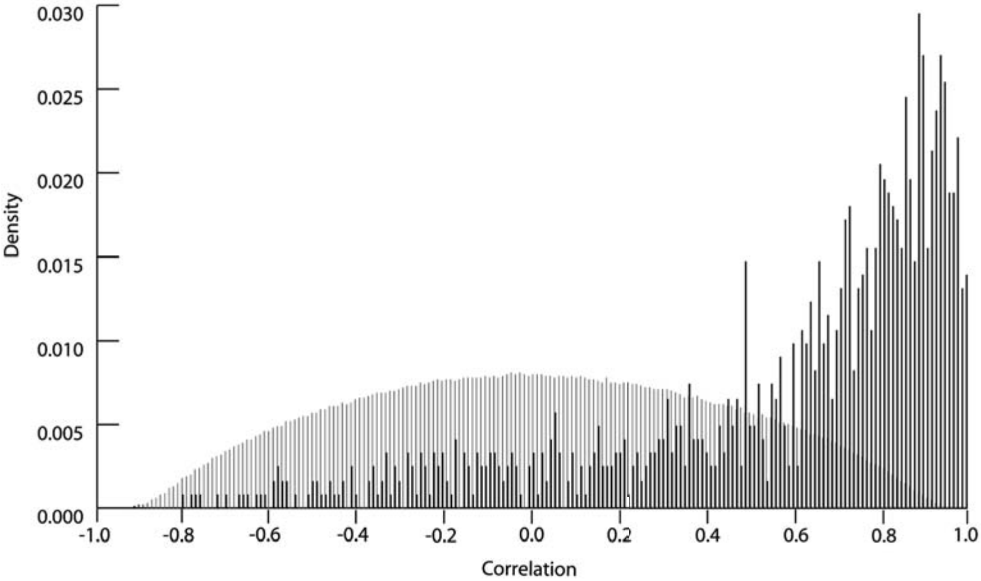


Figure 4: Distribution of expression correlations for gene pairs of *L. plantarum*.
The correlation of gene pairs within a TU (black), shows a clearly different distribution than the correlation of all gene pairs (grey).

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 4 out of 6 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. 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 for experimental verification and support for experimental data verification. Some of the shared and specific motifs will be discussed in orther 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 aminoacyltranferase, the T-box upstream of this gene will have a specific interaction with the methionine tRNA. Moreover, amino-acid specificity of the aminoacyltranferase 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.

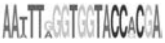
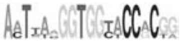
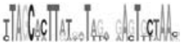
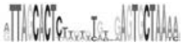
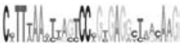
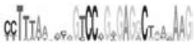
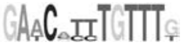
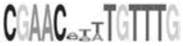
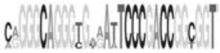
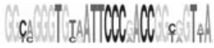
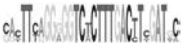
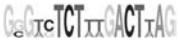
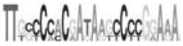
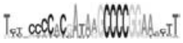
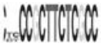
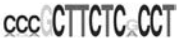
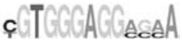
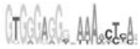
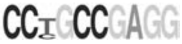
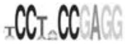
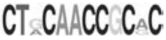
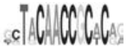
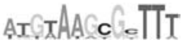
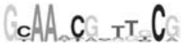
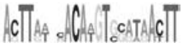
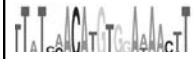
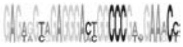
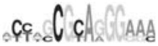
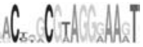
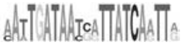
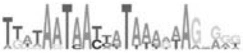
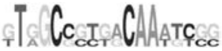
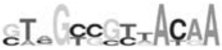
CIRCE

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 *Bacillus subtilis* (27) and *L. lactis* MG1363 (29). Another COT, encoding a metal-dependent membrane-bound protease, often found divergently orientated in front of the transcriptional unit 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 transcriptional units. 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

Table 2 (a color version of the table can be found at page 207): Selected regulatory motifs predicted by both species sets. For a complete list, see the supplementary material. The regulated genes shown are those found in *L. plantarum*; those in other species can be found in supplementary material. 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.

	LAB set	BAC set	Description
1			T-box, element in front of different t-RNA synthetases and related genes.
2			CIRCE element, cis-acting element of HrcA, involved in heat shock response.
3			PyrR element, cis-acting element of PyrR, involved in pyrimidine biosynthesis.
4			LexA element, cis-acting element of DinR, regulator of the SOS regulon.
5			RFN element, DNA element, regulates genes involved in biosynthesis and transport of riboflavin
6			Conserved element in front of peptide release factor 2/B. Regulation of proteins involved in translation.
7			Several different boxes found upstream of genes encoding ribosomal proteins; presumably auto-regulatory sites
			
			
			
			
8			CRE-box, binding site for CcpA, the general catabolite control protein
9			DnaA-box, regulates genes involved in chromosome replication
10	Not identified		S-box, regulates methionine biosynthesis genes in many Gram-positive organisms.
11			THI element, involved in thiamine biosynthesis. Found in different COTs
12	Not identified		FUR, regulates the uptake of Fe
13	Not identified		Stress response
14			Catabolite availability response

n/a = not applicable
1 Genes not in the dataset.
2 If only one TU is found to have the conserved cis-acting element, no correlations for the regulon can be calculated.
3 Since T-boxes have different specificities, depending on their specifier codons, genes regulated by a T-box are not co-regulated.

Regulated genes in L. plantarum	|C|
Many(>10) different t-RNA synthetases. Amino acid biosynthesis and transport genes	n/a[3]
groES-groEL, hrcA-grpE- dnaK, lp0726 (membrane-bound protease)	0.70
pyrR1, pyrP, *pyrR2-pyrAA2-pyrAB2*	0.36
lexA, parC-parE	0.61
ribA-ribB-ribH, lp1887 (transporter)	n/a[1]
prfB	n/a[2]
rplM-rpsI	n/a[2]
infC-rpmI-rplT	n/a[2]
rplK	n/a[2]
rplJ-rplL	n/a[2]
rplU-lp1593-rpmA	n/a[2]
Several PTS systems and other genes involved in sugar metabolism.	0.40
dnaA, dnaN, lp0045	0.70
None	n/a[1]
lp0217-lp0218-lp0219, *thiM-thiD-thiE-lp0116- lp0117*	0.53
None	n/a[1]
lp2521, lp3215, *lp3441 – lp3442, lp3128, lp2807, lp0124, lp0433*	0.64
lp2813, lp3422, lp3553 – araA – araD – araB – araT, lp3591 – rhaD – rhaA – lp3594 – rhaB – lp3596	0.52

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.

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 3 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 *E. 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 can not 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 (Christian Vogl, 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. pentasaceus B. subtilis, S. aureus*), only one of the two (*L. johnsonii, L. brevis, L. delbrueckki, 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 transcriptional unit. 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 4 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 (Table2, 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 1 of the BAC set species, but present in at least 4 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*, *L. acidophilus*) and other lactic acid bacteria (*Leuconostoc mesenteroides* and *Oenococcus oeni*) genomes. All 18 motifs were found to occur in at least 2 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 2, 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 ATPses (CopA, or, in *L. plantarum*, CopA and CopB). *Cop* genes are found in the genus of *Lactobacillaceae* and other lactic acid bacteria, 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 *L. 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 < $1e^{-11}$). Searches in species outside the initial LAB set showed that the motif is present in other lactic acid bacteria. 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 *S. 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 *B. cereus* and *B. anthracis* strains. BLAST searches showed that these genes resemble the *copR* genes of the LAB genomes.

Unknown functions

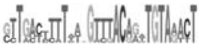
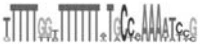
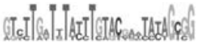
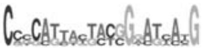
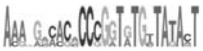
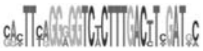
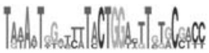
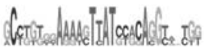
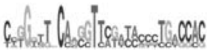
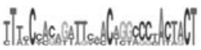
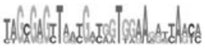
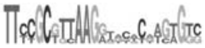
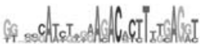
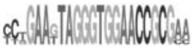
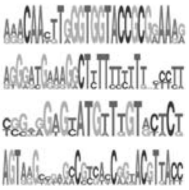
Next to this well-described motif, several LAB-specific motifs were found that seem to be highly conserved among the different LAB species, including *Lactobacillaceae* (or related lactic acid bacteria), but for which the function remains unknown. To the best of our knowledge, none of these motifs have 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 $5.3e^{-11}$), while no hits below threshold (of $1.0e^{-05}$) can be found in the BAC set species. Nevertheless, the genes in the corresponding COT are found to have an orthologue 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.

Table 3 Predicted LAB-specific motifs.
In the fourth column the regulon genes in *L. plantarum* are shown. Genes joined by a hyphen are part of the same TU. A list of regulon genes in other organisms can be found in the supplementary material.

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

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 7 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 preceeding 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 5 and 6 genes), both involved in the breakdown of pentose sugars. Four out of 5 genes (*araADTB*) in the first TU have functions involved in the breakdown of arabinose, while 3 out of 6 genes (*rhaDAB*) in the second TU function in the breakdown of rhamnulose. 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 is 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.

References

1. **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/macroarray analysis (ROMA), and transcriptional profiling approaches. *J Mol Biol, 316, 443-457.*
2. **Conway, T. and Schoolnik, G.K. (2003)** Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. *Mol Microbiol, 47, 879-889.*
3. **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 U S A, 99, 7323-7328.*
4. **Mwangi, M.M. and Siggia, E.D. (2003)** Genome wide identification of regulatory motifs in *Bacillus subtilis*. *BMC Bioinformatics*, **4**, 18.
5. **Thompson, W., Rouchka, E.C. and Lawrence, C.E. (2003)** Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res, 31, 3580-3585.*
6. **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.*
7. **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.*
8. **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., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W. et al. (2003)** Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A, 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 Suppl 1, S329-336.*
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 Database issue, D75-77.*
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 U S A, 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 (Pt 7), 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 procaryotes. 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 U S A, 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.*

CHAPTER 7

Reconstruction of the *Lactobacillus plantarum* WCFS1 regulatory network on basis of correlated gene expression.

Michiel Wels
Lex Overmars
Christof Francke
Michiel Kleerebezem
Roland J. Siezen

Manuscript in preparation

Transcriptome data provide a snapshot of gene expression under a specific experimental condition. In principle, a combination of transcriptome data from many different experiments should enable the elucidation of relations between the activity of certain transcription factors (TFs) and the genes they control, thereby establishing the regulatory network of a cell. To obtain insight in the regulatory network of *Lactobacillus plantarum*, microarray data from more than 70 different conditions were combined and the expression profiles of the transcriptional units (TUs) were compared. The TUs that displayed correlated expression regulation were used to identify putative *cis*-regulatory elements by searching the upstream regions of the TUs for conserved motifs. Predicted regulons were extended by searching for additional motifs in the upstream regions of other TUs with correlated expression regulation. The upstream regions of these TUs were used to refine the putative *cis*-acting elements. In total, *cis*-acting elements were identified for 41 different regulons consisting of at least 4 highly correlated TUs (expression correlation > 0.7). This set of regulons included the well-known regulons of CtsR and LexA (SOS response), but also several novel predicted regulons of coherent functional characteristics. The overall regulatory network reconstructed was visualized using Cytoscape, enabling the study of network interconnectivity. This analysis revealed that the regulatory network of *L. plantarum* is highly interconnected and individual TUs were found to be a member of up to 6 different regulons. In addition, the network contains several subsystems with clear conserved biological functions such as sugar and energy metabolism, nitrogen metabolism and stress response(s).

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 feed-back mechanisms to avoid overreaction to certain stimuli.

The comprehensive interpretation of gene expression data can be greatly enhanced by advanced understanding of the regulatory networks that control their transcription. Such understanding could elucidate regulatory processes underlying specific *in situ* behaviour, for example during gastro-intestinal tract residence or during food fermentation processes, providing targets for optimizing culture performance and improving strain robustness. With the development of large scale post-genomics techniques like genome-wide gene transcription analysis (transcriptomics) and protein binding site analysis (chromatin immunoprecipitation on chip (ChIP-chip) experiments), different efforts to analyze the complete regulatory network of an organism were performed in well-known model organisms such as *E. coli* and *S. cerevisiae* (1-6). Although these first analyses were purely based on experimental data, later efforts included knowledge gathered from literature and databases with curated information on regulatory networks to refine the predicted network (5,7,8). As microarray data provide researchers with a snapshot of the complete transcription profile of the cell, it is an extremely valuable source of information in unraveling regulatory networks. However, individual microarray datasets describe only a co-occurring change in the expression of individual genes, which does not automatically indicate true co-regulation by a common transcription factor (TF). Expression- and regulation-correlation analysis of genes using multiple transcriptome data sets (9) enables the enrichment of truly co-regulated genes.

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 TFs. In these studies, the upstream regions of a group of genes predicted to have the same *cis*-acting element (e.g., on basis of micro-array data) are commonly analyzed using pattern recognition tools such as Gibbs sampling (10) or Expectation Maximization (EM) (11). Using this approach, potential regulons can be identified on basis of shared *cis*-acting elements preceding the co-regulated genes. Combining the knowledge of a shared regulatory binding site (*cis*-acting element) with the correlated change in expression of genes under the applied experimental condition can identify the genes that are truly co-regulated (12-14). Subsequently, conserved motifs in the upstream regions can be used to scan the genome(s) of interest in order to predict the full complement of a regulon. Although these methods have been shown to be valuable in detection of co-regulatory relations in single experiments, large-scale analysis of regulatory networks, using a combination of different transcriptomics experiments, are not yet performed routinely.

In this study we exploited the availability of a large set of microarray data to predict the regulatory network of *L. plantarum*. This large dataset contained microarray experiments with bacteria grown under variable conditions as well as several mutant strains and were related to different scientific interests of the individual researchers involved. *Cis*-acting elements were predicted using the upstream regions of triplets of Transcriptional Units (TUs) with correlated expression. Additional candidate-regulon members were identified on basis of sharing the predicted *cis*-acting element as well as showing correlated expression. In total, 41 sets of co-regulated genes consisting of at least 4 different TUs were identified. These data were visualized as a transcription network using Cytoscape and inspected for overlap of TUs between different regulons. This study shows that correlation analysis of co-regulation in multiple transcriptome datasets combined with *cis*-acing element prediction provides a valuable strategy for the prediction of regulons and regulatory networks.

Materials and Methods:

A flowchart of the followed procedure is depicted in Figure 1.

Expression data

Expression data was obtained from 72 microarray experiments of *L. plantarum* WCFS1 using Agilent oligo-based arrays. The tested experimental conditions were highly variable and range from stress conditions (such as ethanol and peroxide challenge) to knockout or overexpression of specific (metabolic) genes (GEO accession codes GSM136883 – 136888, GSM206844 – GSM206852, GSM 215123 – GSM215128, GSM217127 – GSM217132 and GSM217146 – GSM217149). All array measurements were normalized by local fitting of an M-A plot using the implementation of the LOWESS algorithm in R (http://www.r-project.org).

TU prediction:

TUs were predicted based on the methods described by (15) except that the intergenic distances were taken as <100 nt instead of

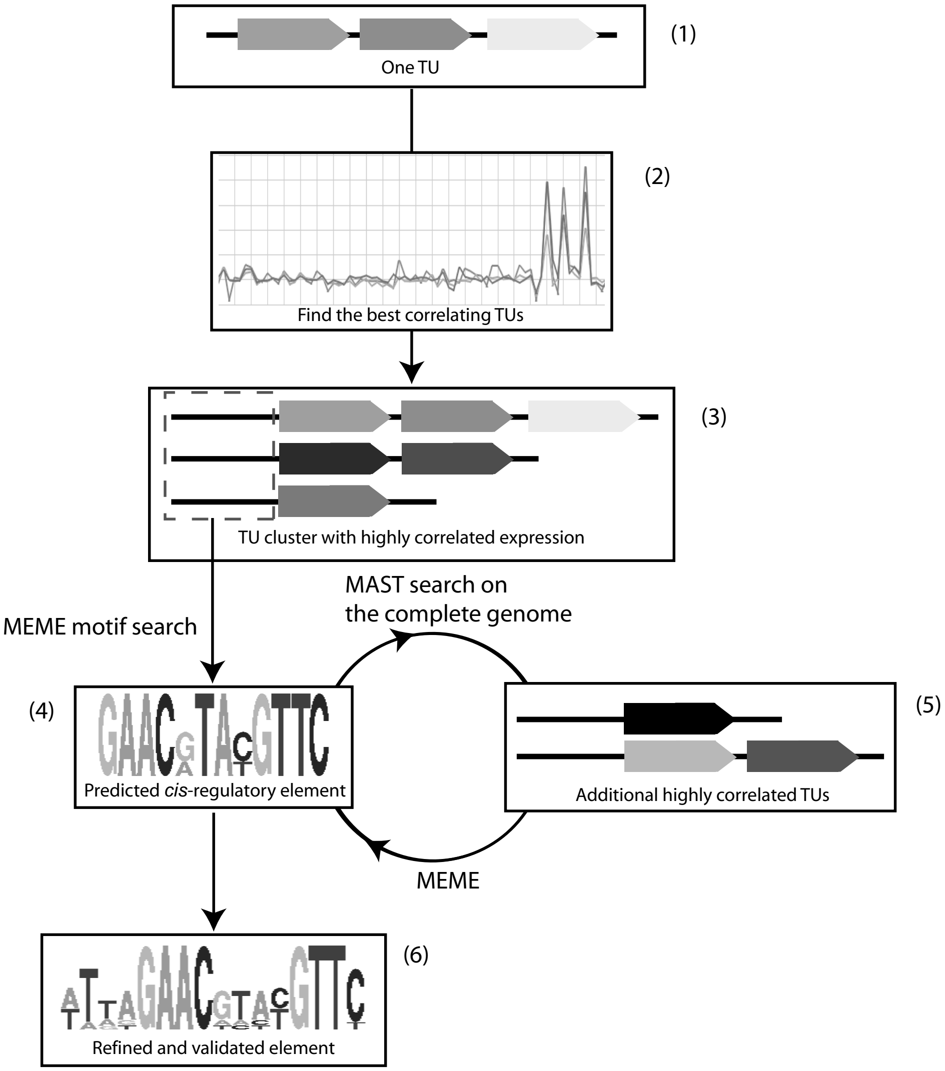

(1)
One TU
(2)
Find the best correlating TUs
(3)
TU cluster with highly correlated expression
MEME motif search
MAST search on the complete genome
(4)
GAACGTACGTTC
Predicted *cis*-regulatory element
(5)
Additional highly correlated TUs
MEME
(6)
ATTAGAACGTACGTTC
Refined and validated element

Figure 1 (a color version of the figure can be found at page 180): Flowchart of the followed procedure.
1= query TU, 2= identification of co-regulated TUs in a large set of experiments, 3= selection of upstream sequences of co-regulated TU cluster, 4= shared cis-element detection, 5= iteration procedure to expand or complete regulon, 6= refined regulon cis-element

<50 nt to decrease the likelihood of splitting one functional TU into multiple, smaller TUs. Although this increase in intergenic distance can decrease the number of TUs with correlated expression, TUs that are inappropriately divided into multiple smaller TUs will drastically increase the level of noise in the motif prediction as they are bound to display highly correlated expression without sharing an additional TF binding site in their upstream sequences.

Correlation analysis:

Pearson correlation of gene expression was calculated between TUs on basis of gene pairs. Only genes that showed a change in ^{2}log expression ratio of at least 1 (or -1) in at least one of the performed experiments were considered in the analysis. Correlations between TUs were predicted by calculating the mean of all Pearson correlations for all possible gene pairs between the different TUs. These TU-TU correlations were stored, together with the gene correlations in a MySQL database (http://www.mysql.com) for fast and easy access. For every TU for which microarray data was present, the two best correlating operons were extracted from the database. The dataset was filtered for low scoring triplets (individual TU-TU correlation < 0.7) and triplet redundancy (the same triplets, resulting from a different TU as a starting point) before applying motif prediction. The chosen correlation value of 0.7 was found to be significant with a confidence interval of 95%.

Motif prediction:

MEME software (11) was used to predict regulatory motifs from upstream regions of 300 nt per TU. MEME settings applied were: minimum length 5 nt, maximum length 20 nt, 4 different motifs, found in at least 2 out of 3 sequences and find motifs on both strands (-revcomp). Other settings were set at default. All predicted motifs were used in a MAST search (16) against the upstream regions of all predicted TUs. If hits were found that were not part of the original regulon triplet, their correlation with the starting triplet was calculated. All hits with a correlation higher than 0.7 with the original triplet were used to build a new, more specific motif with MEME. This procedure was iterated for all motifs until no additional, well-correlating hits were found.

Regulon filtering:

All iterated motifs were used for a final MAST search in all TU upstream regions (cutoff $1e^{-07}$), leading to a final set of putative co-regulated TUs (regulon). Expression correlation for individual TUs with the regulon were calculated by the mean correlation of the subject TU with all TUs used in the final MEME iteration. On basis of this analysis, TUs with a *cis*-acting element in the upstream region were classified in two categories; truely co-regulated (correlation >0.7) and putative false-positive (correlation <0.7). In some cases, no categorization could be made due to lack of microarray data for that specific TU.

Functional classification of TUs and regulons

Functional (sub-) classifications of the annotation of the genes within TUs were compared between TUs in a regulon. If a certain (sub-)class was found in more than one TU of a regulon, the genes in this regulon were manually inspected for functional coherence. The sub classes “Not conserved: other”, “Conserved: other”, “Conserved: putative function” within the main class “Hypothetical proteins” and the sub class “Unknown substrate” within the main class “Transport and Binding” were not considered to be relevant, since the genes within these categories do not share a defined functional relation. This analysis was performed on all initial TU triplets as well as the iterated regulons.

Network visualization

All regulons were displayed using Cytoscape (http://www.cytoscape.org). TUs were displayed as nodes and connected by edges if they were present in the same regulon. Different regulons were manually assigned different colors. Nodes describing a TU that encoded for a regulatory protein were manually colored green.

Results:

To predict the regulatory network of *L. plantarum*, the first step is to predict *cis*-regulatory elements on basis of the upstream sequences of TUs with correlated expression. As a starting point, the upstream sequences of triplets of highly correlated TUs were selected. Three TUs were regarded as the minimum to distinguish noise from biological overrepresentation. For 523 out of 1735 TUs at least one gene showed significant elevated expression in at least one of the microarray experiments (Table 1). Out of these 523 TUs, 345 appeared suitable for triplet formation since two additional operons could be identified that displayed correlated expression modulation above the threshold (0.7) (Table 1). These initial 345 triplets were filtered for redundancy, reducing the number of triplets to 286. These candidate regulons were searched for functional coherence between the TUs on basis of the functional classification of the genes in the TU, as described in the original *L. plantarum* genome annotation (17). In total, 45 triplets were found to contain at least two TUs sharing the same functional (sub-) class; in 9 cases all three TUs encompassed at least a single gene sharing the same functional subclass. However, in several cases (4 candidate TU-triplets) the shared functional class was “Transport and binding proteins”, which is relatively uninformative, since in these cases the transporter genes in the triplets were not annotated at the level of substrate specificity. Only one of these candidate TU-triplets displayed similarity in the metabolite specificity (Amino acid transport). The other 5 triplets encompassed a more stringent functional parallel (Table 2). In one candidate triplet all three TUs encoded different prophage functions and were genetically closely linked on the chromosome. This observation of cassette-like organization and conserved regulation of prophage genes is in agreement with previous observations (for an overview, see (18)). In two other cases, all three TUs encoded genes involved in plantaricin production and immunity (19). In these two cases, the TU triplets only partially overlapped with each other and the predicted motifs of the triplets were not alike. MAST searches did not gain extra hits in the upstream regions of other plantaricin related TUs. The detection of several of these functionally conserved TU triplets is a good indication that the TUs with the highest correlated expression are functionally related and are likely to share regulatory characteristics controlled by the same TF.

The upstream sequences of all 286 triplets were subject to MEME (11) analysis to predict conserved *cis*-regulatory elements. Subsequent MAST (16) analysis in the upstream regions of all TUs showed that with 62 of the predicted motifs at least one additional homologous motif was identified that preceded a TU displaying significant expression correlation with the initial TU triplet (>0.7). In 5 cases two different motifs were identified from the same triplet of TUs. The upstream regions of these additional TUs were used, in combination with those of the original triplet, to predict refined MEME motifs. This procedure was iterated for all motifs until saturation, which was defined

Table 1: Overview of TU and regulon prediction

Total Number of TUs	**1735**
Genes on the microarray	3078
Genes with elevated expression	802
TUs with elevated expression	523
TUs present in triplets	345
Triplets of TUs	286
Regulons of size > 4	31
TUs in regulons	112
Genes in regulons	225

Table 2: TU triplets (regulons) sharing a functional class.
TU triplets sharing the class “Transport and binding proteins” are not shown.

Triplet	TU	Gene name	Function
Triplet_884	TU_884	*lp_1903 (clpB)*	ATP-dependent Clp protease, ATP-binding subunit ClpB
	TU_601	*lp_1269 (clpE)*	ATP-dependent Clp protease, ATP-binding subunit ClpE
	TU_366	*lp_0786 (clpP)*	endopeptidase Clp, proteolytic subunit
Triplet_527	TU_522	*lp_1101 (ldhL2)*	L-lactate dehydrogenase
		lp_1102	cation transport protein
	TU_528	*lp_1113*	fumarate reductase, flavoprotein subunit precursor
	TU_527	*lp_1112 (fum)*	fumarate hydratase
Triplet_1119	TU_1121	*lp_2439*	prophage P2a protein 18
		lp_2440	prophage P2a protein 17
		lp_2441	prophage P2a protein 16
		lp_2442	prophage P2a protein 15
	TU_1120	*lp_2432*	prophage P2a protein 25
		lp_2433	prophage P2a protein 24
		lp_2434	prophage P2a protein 23
		lp_2435	prophage P2a protein 22
		lp_2436	prophage P2a protein 21
		lp_2437	prophage P2a protein 20
	TU_1119	*lp_2426*	prophage P2a protein 31
		lp_2427	prophage P2a protein 30
		lp_2428	prophage P2a protein 29
		lp_2429	prophage P2a protein 28
		lp_2430	prophage P2a protein 27
		lp_2431	prophage P2a protein 26
Triplet_209	TU_211	*lp_0411 (plnO)*	plantaricin biosynthesis protein PlnO
		lp_0412 (plnP)	immunity protein PlnP, membrane-bound protease CAAX family
	TU_210	*lp_0410 (plnN)*	bacteriocin precursor peptide PlnN (putative)
	TU_209	*lp_0409 (plnM)*	immunity protein PlnM
Triplet_213	TU_211	*lp_0411 (plnO)*	plantaricin biosynthesis protein PlnO
		lp_0412 (plnP)	immunity protein PlnP, membrane-bound protease CAAX family
	TU_210	*lp_0410 (plnN)*	bacteriocin precursor peptide PlnN (putative)
	TU_213	*lp_0413 (plnQ)*	plantaricin biosynthesis protein PlnQ

Main class	**Subclass**
Cellular processes	Adaptations and atypical conditions (Clp system)
Cellular processes	Adaptations and atypical conditions (Clp system)
Cellular processes	Adaptations and atypical conditions (Clp system)
Energy metabolism	Fermentation
Transport and binding proteins	Cations
Energy metabolism	Electron transport
Energy metabolism	TCA cycle
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Other categories	Phage and prophage related functions
Cellular processes	Toxin production and resistance (plantaricin)
Cellular processes	Toxin production and resistance (plantaricin)
Cellular processes	Toxin production and resistance (plantaricin)
Cellular processes	Toxin production and resistance (plantaricin)
Cellular processes	Toxin production and resistance (plantaricin)
Cellular processes	Toxin production and resistance (plantaricin)
Cellular processes	Toxin production and resistance (plantaricin)
Cellular processes	Toxin production and resistance (plantaricin)

by lack of discovery of additional TUs with significant correlation to the query candidate regulon. Complete saturation for all initial candidate regulon triplets was reached after 7 iteration steps. The complete set of regulons was checked for redundancy and duplicates were cleared from the set, resulting in a final prediction of 50 different motifs predicted from the upstream regions of more than three different TUs. Finally, MAST searches were performed for all these motifs to identify all occurrences in the genome sequence. On basis of the MEME analysis, all TUs that showed to be highly correlated in expression (>0.7) and shared an upstream motif with a p-value < $1.0e^{-07}$ were regarded as regulon members. A TU size distribution graph was made for the 50 different motifs (Figure 2). Although the iteration procedure was based on finding additional members associated with the original triplets and thus generate regulons of at least four TU members, nineteen regulons were identified of a size smaller than 4 TUs. The MEME analysis of these nineteen regulons was shown to be based on sequences that eventually scored worse than the final, more stringent MAST cutoff. These regulons were not further investigated in this study. The largest regulon that was identified in *L. plantarum* consisted of 11 different TUs, encompassing a total of 19 genes. In addition, almost 85% of all validated regulons has a size of 6 TUs or less. This observation is in agreement with the commonly accepted notion that only a limited number of globally acting regulators exist in bacteria, which are expected to regulate expression of multiple TUs (20). Although it is likely that larger and more globally acting regulons exist in *L. plantarum*, these were not detected in this analysis. The lack of detection of these regulons could be the result of the common property of these globally acting TFs that the level of control of gene expression is part of a hierarchial regulatory system that includes fine-tuning of subsets of TUs within their regulon by more locally acting TFs, as shown for CcpA and CcpB in *B. subtilis* (21) and recently suggested to occur for CcpA with many other LacI family members (22). All identified regulons of >3 TUs can be found at http://www.cmbi.ru.nl/~mwels/Thesis/Chapter_7.

Several of the predicted regulons harboured genes with a clearly coherent biological role and represent known regulons in various bacteria. Two examples are discussed below.

The CtsR regulon

One of the iteratively expanded regulons contained four stress-related TUs that displayed strongly concerted regulation of expression. Three of these TUs encode proteins involved in the Clp protease complex, while the fourth TU contains a single gene annotated as small heat shock protein (*hsp1*). Upstream of one of these TUs (TU_884) two copies of the corresponding motif were identified (p-values $4e^{-07}$ and $5e^{-10}$). The predicted *cis*-element of this regulon is a perfect direct repeat (AAGGTCA-(N3)-AAGGTCA) and strongly resembles the consensus binding site of CtsR (GGTCAAANANGGTCAAA) as described for *B. subtilis* in (23). CtsR is a stress-response regulator known to be involved in Clp activation in several different *Firmicutes* e.g. *B. subtilis* (24), *Staphylococcus aureus* (25), *Streptococcus pneumoniae* (26) and *Lactococcus lactis* (27). The involvement of CtsR in regulation of this regulon could be confirmed, since the transcriptome database encompasses experiments in which expression profiles of wild-type and *ctsR*-mutant strains are compared; these experiments displayed the highest gene expression ratios of the TUs included in this regulon. Moreover, the predicted CtsR-regulon of *L. plantarum* encompasses many genes orthologous to

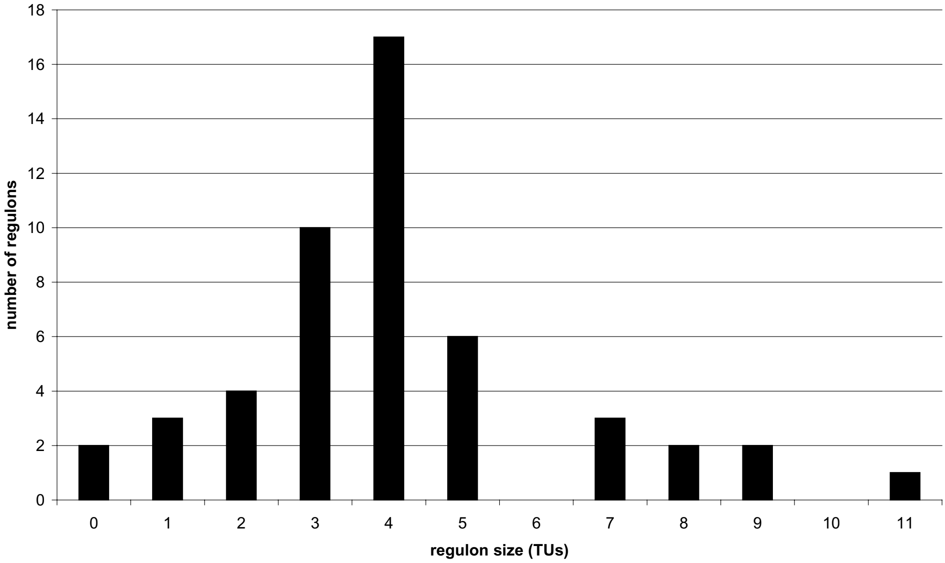


Figure 2: Regulon size distribution of iteratively expanded regulons
Regulons of a size < 3 were not analyzed further (see methods)

CtsR-regulated genes in *B. subtilis,* including c*lpE* and c*lpB* (28). However, the genome of *B. subtilis* appears to lack a gene encoding an ortholog of the *L. plantarum hsp1* gene, suggesting that the CtsR regulons of *B. subtilis* and *L. plantarum* do not control identical sets of genes. Moreover, the CtsR regulon identified in *L. plantarum* does not include a predicted autoregulatory circuit for the control of expression of the *ctsR* gene itself, which is in clear contrast to the situation found in *B. subtilis* (29).

The SOS-regulon (LexA)

Two of the largest predicted regulons (both encompassing 9 TUs) contained many DNA-damage repair functions. As the predicted *cis*-acting elements of these two regulons are very alike and 7 out of the 9 TUs appeared to be shared between the two regulons, they were merged into a single regulon consisting of 11 TU members, encompassing in total 16 genes. These 16 genes belonged to several functional classes, including "DNA metabolism" (6 genes), "Transcription", "Protein synthesis", and "Regulatory functions" (all 1 gene), while the residual genes belonged to the category of hypothetical proteins but with putative functions as segregation helicase (*lp_1543*) and exopolyphosphatase-related protein (l*p_2279*). The regulon motif identified was a highly conserved palindromic sequence (GAAC-(N4)-GTTC), resembling the binding site of LexA (or DinR) involved in regulation of SOS response in various organisms. The SOS regulon is found in almost all lineages of bacteria. It responds to potential DNA-damaging agents and encodes genes involved in repairing DNA damage (for a recent review, see (30)). Notably, LexA itself (*lp_2063,* the only gene classified as "Regulatory functions") also appeared to be a member of the regulon. The SOS regulon has been described in the past in several different organisms, including *E. coli* and *B. subtilis* (31,32).

The regulatory process of DNA-damage repair is based upon cleavage of LexA by RecA, a protein activated by binding to single-stranded DNA. Interestingely, *recA* was not found to be part of the SOS-regulon in *L. plantarum*. This observation is in contrast to the regulon organization in other bacteria (30) and probably caused by the increased length of the intergenic region that we used in TU prediction (from 50 to 100 nt). Analysis of the 91-bp intergenic region in between *recA* and *cinA* showed that there is a conserved LexA-binding site in the upstream sequence of *recA*, connecting this locus to the SOS locus. Detailed analysis showed that *recA* displayed highly correlated expression with several members in the SOS regulon (e.g. *umuC*: 0.82; *lexA*: 0.75; *dinP*: 0.88), while *recA* and *cinA* had a low correlation in expression (0.50).

In addition to the 11 *cis*-acting elements in the upstream regions of TUs with a high correlated expression, 18 additional occurrences of this motif were found. In these 18 cases, the downstream located TUs had a correlated expression below 0.7 with the predicted SOS-regulon. In three cases, these TUs had a correlated expression with the regulon between 0.5 and 0.7 and two of these TUs encoded functions related to DNA metabolism (Table 3), while the third contained an integrase-coding gene that is most likely of prophage origin as based on genome context information. Of the remaining 15 identified motifs, only one was found in the upstream region of a TU for which microarray data were available. However, this TU displayed no expression correlation (-0.09) with the identified members of the regulon. In 5 cases, a single motif was identified that could putatively control regulation of two different (divergently orientated) TUs. Only in a single case (TU_1472 and TU_1473) actual correlation of expression was observed suggesting that these TUs share this regulatory motif.

The *L. plantarum* regulon prediction and analysis was used to perform cross-species regulon network comparison to *B. subtilis*. Initial analyses immediately showed a limited number of shared genes within common regulons. As an example, of the 16 genes found in the predicted *L. plantarum* SOS-regulon only 11 appeared to be conserved in *B. subtilis*. Moreover, only 4 of these genes have been reported to belong to the LexA regulon in *B. subtilis* (based on the information present in the DBTBS (33)). Of other putative regulon members with a LexA-binding site in *L. plantarum* (Table 3), 5 orthologs were identified in *B. subtilis* of which 4 are subject to LexA regulation in that host (*lp_1612 (gmk1)* , *lp_1839* (*parC)*, *lp_1840* (*parE)* and *lp_2062*). It seems likely that these genes are also part of the LexA regulon in *L. plantarum*, but could possibly be subject to additional transcriptional control mechanisms in this host. The relatively small overlap between the regulons of *B. subtilis* and *L. plantarum* is another example of the large differences in SOS-response in different organisms as was already apparent from the major differences between the *B. subtilis* and *E. coli* SOS regulons (30).

Reconstruction and analysis of the *L. plantarum* regulatory network

In addition to the functional coherence of genes within the different regulons, it was apparent that several combinations of regulons displayed functional coherence as well. Analogous to what was observed with the final SOS regulon that was composed of two initially identified regulons (see above), many genes were part of more than a single regulon and hence their upstream regions contain multiple, different regulatory motifs. As an example, TU_1398 (*lp_3009 – lp_3011*, coding for two different cellobiose-PTS subunits and a 6-phospho-beta-glucosidase)) was found to be part of 5 different regulons. On basis of these overlapping regulons and shared genes, it was possible to reconstruct an initial regulatory network of *L. plantarum* WCFS1 (Figure 3).

Analysis of the regulatory network disclosed some remarkable characteristics. Firstly, the overall network can be divided into eight smaller subsystems, some of which consist of a single or isolated regulon (like the CtsR regulon), while others form a highly intertwined and interacting network, sharing interactions between up to 10 regulons in a single connected regulatory network. The network interactions vary in degree of connectivity, ranging from regulon pairs that share only a single TU, to pairs of regulons that contain up to 7 common TUs. This dense interconnected organization is in line with the Dense Overlapping Regulon (DOR) structures found in the analysis of the regulatory network of *E. coli* (6).The two largest regulatory network subsystems identified in *L. plantarum* encompass 8 and 10 regulons. The functional coherence of genes encompassed within the identified regulatory network subsystem was analyzed in more detail. These analyses showed that several of the subsystems, including the two largest subsystems (Figure 3, http://www.cmbi.ru.nl/~mwels/Thesis/Chapter_7), displayed a large degree of functional coherence in terms of functional annotation of the genes contained within the regulatory network subsystem.

The largest subsystem contained 5 TUs that are highly connected (present in 3 different regulons). Three of these TUs encoded genes involved in the biosynthesis of amino acids, *i.e.* TU_266 *(lp_0526, lp_0527)* and TU_267 *(lp_0528 – lp_0530)* encode genes involved in glutamate biosynthesis, while TU_1655 (*lp_3497 – lp_3499*) encodes genes responsible for synthesis of aromatic amino acids. The two residual TUs contained genes that potentially play a role in amino acid metabolism: TU_1492 contains a gene codon for a transporters with unknown specificity (*lp_3183*) and the other (TU_64) contains two hypothetical proteins (*lp_0130 – lp_0131*), of which one (*lp_0131*) is predicted to be located in the cell membrane.

Regulation network of sugar and energy metabolism

The second largest regulatory network subsystem contains 8 connected regulons, encompassing 25 TUs containing a total of 61 genes (Table 4). Many of these genes encode sugar metabolism related functions, including 8 genes encoding different PTS subunits, 3 genes encoding sugar-metabolism related regulatory proteins and 9 polysaccharide or sugar-metabolism related enzymes. In addition, this regulatory network subsystem included several TUs containing genes involved in energy metabolism, including an oxidoreductase, a transaldolase, a phosphoglycolate phosphatase and a phosphoglycerate mutase. These sugar- and energy-metabolism related TUs were highly interconnected and present in multiple different regulons (Table 4). Next to these, additional genes were encompassed in this network subsystem that could potentially be related to sugar metabolism: 3 transporters with unknown substrate specificity and 4 genes encoding cell-envelope proteins. In contrast, 16 genes could not be classified in sugar metabolism. These genes were organized in single TUs. For the remaining genes it was unclear to decide upon their relation to sugar and energy metabolism as they lacked a specific annotation.

Table 3: SOS regulon of *L. plantarum*
All regulon members are listed (TUs and encoded genes). P-values of MAST hits (motifs) in upstream regions are shown, together with correlation values of gene expression.
* hit found in the intergenic region between two divergently transcribed TUs, possible false-positive
n/a no expression data available for the TU

TU	Gene	Function	Main class	P-value	Gene expression Correlation
True regulon members (correlation > 0.70)					
1473*	*lp_3142*	Unknown	Hypothetical proteins	2.1e-11	0.87
1409	*lp_3022*	Unknown	Hypothetical proteins	2.0e-08	0.86
	lp_3023 (umuC)	UV-damage repair protein	DNA metabolism		
1472*	*lp_3141*	Unknown	Hypothetical proteins	2.1e-11	0.85
724	*lp_1543 (cshA2)*	1 segregation helicase (putative)	Hypothetical proteins	8.2e-10	0.84
1240	*lp_2693 (rexA)*	ATP-dependent nuclease, subunit A	DNA metabolism	2.1e-08	0.83
	lp_2694 (rexB)	ATP-dependent nuclease, subunit A	DNA metabolism		
755*	*lp_1611*	Unknown	Hypothetical proteins	7.8e-10	0.83
1062	*lp_2278 (rhe3)*	ATP-dependent RNA helicase	Transcription	2.7e-10	0.79
	lp_2279	Exopolyphosphatase-related protein (putative)	Hypothetical proteins		
	lp_2280 (dinP)	DNA-damage-inducible protein P	DNA metabolism		
965*	*lp_2063 (lexA)*	transcription repressor of the SOS regulon	Regulatory functions	4.2e-08	0.79
1064*	*lp_2285 (queA)*	S-adenosylmethionine tRNA ribosyltransferase-isomerase	Protein synthesis	1.0e-08	0.73
	lp_2286 (ruvB)	holliday junction DNA helicase RuvB	DNA metabolism		
	lp_2287 (ruvA)	holliday junction DNA helicase RuvA	DNA metabolism		
71*	*lp_0145*	Unknown	Hypothetical proteins	2.8e-08	0.70
Additional putative members (correlation between 0.70 and 0.50)					
158	*lp_0305 (gcsH1)*	glycine cleavage system, H protein	Energy Metabolism	2.1e-08	0.68
	lp_0306	Unknown	Hypothetical proteins		
	lp_0307	Unknown	Hypothetical proteins		
	lp_0308	DNA Helicase	DNA metabolism		
310	*lp_0624*	prophage P1 protein 1, integrase	Other categories	3.1e-07	0.62
858	*lp_1839 (parC)*	topoisomerase IV, subunit A	DNA metabolism	1.6e-07	0.58
	lp_1840 (parE)	topoisomerase IV, subunit B	DNA metabolism		
Additional non-correlated members					
756*	*lp_1612 (gmk1)*	guanylate kinase	Purines, pyrimidines, nucleosides and nucleotides	7.8e-10	-0.09
	lp_1613 (rpoZ)	DNA-directed RNA polymerase, omega subunit	Transcription		
Not defined					
469	*lp_0961*	Unknown	Hypothetical proteins	2.7e-10	n/a

1065*	*lp_2289*	Unknown	Hypothetical proteins	1.0e-08	n/a
1144	*lp_2504*	Unknown	Hypothetical proteins	4.3e-08	n/a
636	*lp_1346 (asd1)*	aspartate-semialdehyde dehydrogenase	Amino acid biosynthesis	3.0e-08	n/a
305	*lp_2830 (ansB)*	aspartate ammonia-lyase	Amino acid biosynthesis	3.8e-07	n/a
70*	*lp_0141*	extracellular protein	Cell envelope	2.8e-08	n/a
331	*lp_0709 (galE1)*	UDP-glucose 4-epimerase	Purines, pyrimidines, nucleosides and nucleotides	1.0e-08	n/a
938	*lp_1997*	integrase, fragment	Other categories	4.5e-08	n/a
964*	*lp_2062*	Unknown	Hypothetical proteins	4.2e-08	n/a
572	*lp_1215 (cps3A)*	Glycosyltransferase	Cell envelope	3.3e-07	n/a
	lp_1216 (cps3B)	Glycosyltransferase	Cell envelope		
469	*lp_0961*	Unknown	Hypothetical proteins	1.4e-07	n/a

Analysis of the predicted *cis*-acting elements within this sugar-related regulatory network revealed that there is only a limited overlap in motifs. Only two predicted motifs appeared to be in the same position within all three overlapping TUs (TU_434, TU_1665 and TU_1705). Consequently, the consensus sequences of these motifs are identical (GAAAACGCTATC). This consensus sequence resembles the consensus sequence of the known catabolite repression element (*cre*) for Gram-positive organisms (WTGNAANCGNWNNCW). *Cre* is known to be recognized by CcpA but is also shown to bind different members of the LacI-family of regulators (22). The regulators of this family are known to be involved in the regulation of many different sugar metabolism genes (chapter 2 of this thesis, (22)). In addition to the two *cre*-like occurrences, only one additional case of overlap was observed between the predicted motifs. In the upstream region of TU_1674 two different motifs were found to partially overlap. This was not immediately apparent from the motif logos, as the motifs were identified on two different strands. Next to TU_1674 we found one other TU (1351) to also share these two *cis*-acting elements, but the motifs were not found to spatially overlap, suggesting that they are not two representations of a single motif.

Discussion:

Analysis of a large set of microarray experiments obtained from a variety of experimental conditions and mutant-derivative strains showed to be of great value for obtaining data-driven insight in the regulatory network of *L. plantarum*. Correlated expression data could be used to identify co-regulated TUs. Combined with detection of shared *cis*-acting regulatory elements, a regulon prediction database could be constructed that allowed reconstruction of the regulatory network of *L. plantarum* WCFS1. In many cases the *cis*-acting elements were not previously identified. Moreover, functional coherence within regulons and regulatory network subunits became apparent, supporting the biological relevance of the network constructed.

The two largest subsystems within the regulatory network seem to represent carbon (sugars) and nitrogen (amino acid) metabolism. As the global regulatory protein for nitrogen regulation is not found in *L. plantarum* (CodY, the master regulator in nitrogen regulation in most *Firmicutes* was found to be absent

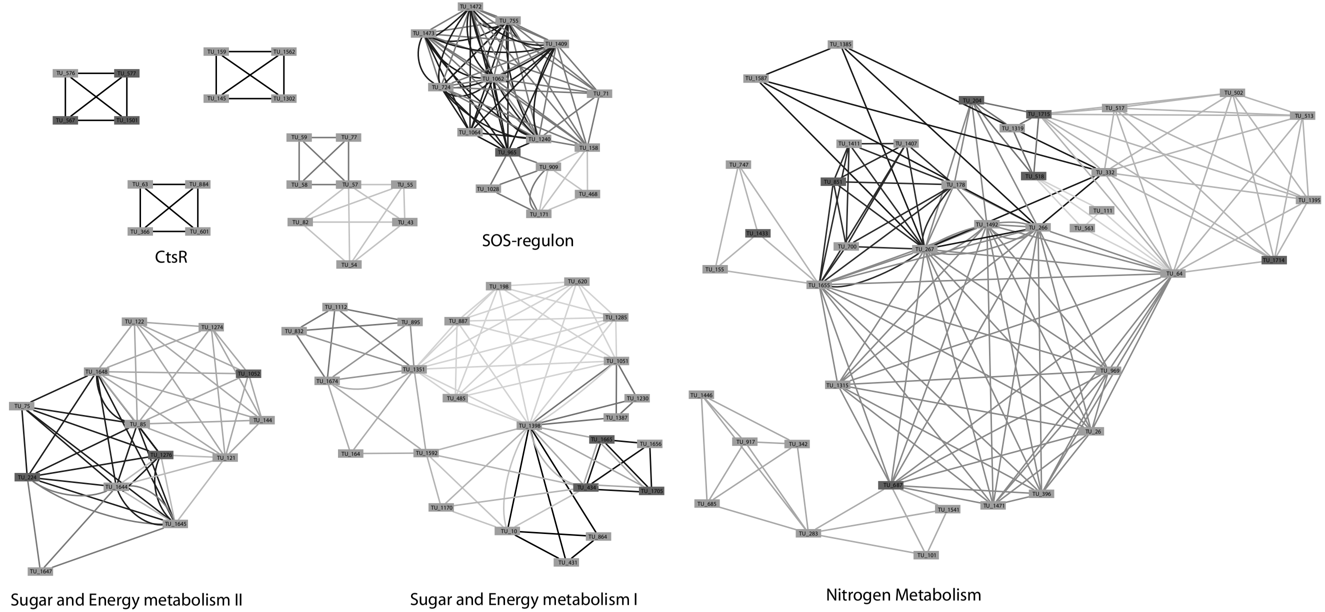
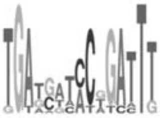
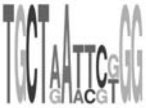
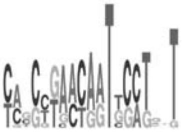
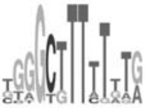
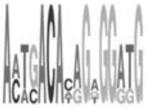
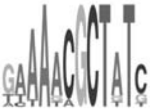
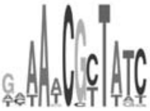
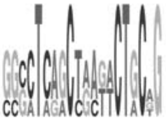


Figure 3 (a color version of the figure can be found at page 181): Regulatory network of *L. plantarum* based on iteratively expanded regulons.
Nodes represent different TUs; lines connect TUs that are member of the same regulon, distinguished by different colors per regulon. TUs colored in green contain at least one gene encoding a regulatory protein. A detailed view on the network can be found at www.cmbi.ru.nl/~mwels/Chapter_7.

Table 4: Regulons clustered in the sugar metabolism regulatory network subsystem.
TUs in bold are shared among multiple different regulons. 1 LacI family motifs. More details on these proposed regulons can be found at http://www.cmbi.ru.nl/~mwels/Thesis/Chapter_7.

Motif	TUs	Main classes
	TU_832 TU_895 TU_1112 TU_1351 TU_1674	Energy metabolism (glycolysis) Transport Cell envelope (lipoproteins)/Transport Hypothetical proteins Central intermediary metabolism
	TU_164 TU_1351 TU_1592 TU_1674	Hypothetical proteins Hypothetical proteins Other categories (prophage-related) Central intermediary metabolism
	TU_198 TU_485 TU_620 TU_887 TU_1285 TU_1051 TU_1351 TU_1398	Hypothetical proteins Hypothetical proteins Cell envelope (Cell surface) Transport (Multidrug) Cell envelope (LPXTG) Hypothetical proteins Hypothetical proteins Transport (PTS)/Energy metabolism (Sugars)
	TU_1170 TU_10 TU_434 TU_1398 TU_1592	Amino acid biosynthesis (Histidine) Central intermediary metabolism (Polysaccharides) Regulatory functions (BglB)/ Transport (PTS)/ Energy metabolism (Sugars) Transport (PTS)/Energy metabolism (Sugars) Other categories (prophage-related)
	TU_431 TU_864 TU_10 TU_1398	Cellular processes (Chaperones) Hypothetical proteins Central intermediary metabolism (Polysaccharides) Transport (PTS)/Energy metabolism (Sugars)
 1	TU_434 TU_1398 TU_1665 TU_1705	Regulatory functions (BglB)/ Transport (PTS)/ Energy metabolism (Sugars) Transport (PTS)/Energy metabolism (Sugars) Regulatory functions/ Transport (PTS)/ Hypothetical proteins Regulatory functions/ Transport (PTS)/ Energy metabolism (Sugars and General)
 1	TU_1656 TU_434 TU_1665 TU_1705	Energy metabolism/ Hypothetical proteins Regulatory functions (BglB)/ Transport (PTS)/ Energy metabolism (Sugars) Regulatory functions/ Transport (PTS)/ Hypothetical proteins Regulatory functions/ Transport (PTS)/ Energy metabolism (Sugars and General)
	TU_1230 TU_1387 TU_1051 TU_1398	Transport Cell envelope (Teichoic acid biosynthesis) Hypothetical proteins Transport (PTS)/Energy metabolism (Sugars)

in *L. plantarum* (20)) it is interesting to see if candidate regulators are present in this nitrogen metabolism subsystem. Analysis of the subsystem revealed the presence of eight genes encoding regulator functions (*lp_0396, lp_0889, lp_1092, lp_1443, lp_1821, lp_3079, lp_3649* and *lp_3650*). A BLAST search showed that none of these regulatory proteins share sequence similarity with the CodY protein sequence from *B. subtilis* (e-value cutoff of 1). However, this observation does not exclude the possibility that one of these regulatory proteins acts as the global regulator of nitrogen metabolism in *L. plantarum*.

Although these results show that correlated expression over a large set of microarray data can be of great help in unraveling the regulatory network of an organism, the amount and source of microarray data probably has a great influence on the usefulness of the data. Although this analysis incorporated the data of over 70 microarrays, only 802 genes (<25%) displayed sufficient levels of differential expression in any of the experiments to be included in the co-regulation analysis performed here. An increase in the number of experiments should result in more variability in the expression pattern of an individual gene. This variability will result in the incorporation of a higher number of genes and TUs in the initial analysis and thus increase the scope and resolution of the network. Eventually, an increase in the number of regulons could potentially link more different subsystems to each other and lead to one large, highly interconnected regulatory network. Nevertheless, the relatively small number of genes incorporated in the analysis performed here still resulted in the identification of a reasonable number of regulons (31). As our method is fully automatic, the regulatory network can be easily fine-tuned and updated when the amount of transcriptomics data increases.

Regulatory networks can be of great help in the analysis of individual transcriptomics data sets (1-3). In most of the studies, interactions between TF and regulated TUs (or genes) are predicted based on the full complement of transcriptomics data and then used to analyze the genetic response of a cell to the imposed changes of a specific experiment. The information gathered from these data is then used to find new TF targets and reconstruct the status of a regulatory network under a specific condition.

Although in some cases the TF binding to a *cis*-acting element can be predicted on basis of literature data (e.g. SOS response, regulation of Clp proteases), a major drawback of this study is the lack of data that links a *cis*-acting element to a transcription factor. TF-binding data from for example ChIP-chip experiments will lead to the prediction of a regulatory network with a higher resolution. Although we observe a degree of connectivity that is comparable to earlier observations in other bacteria, detailed information on TF-*cis*-regulatory element interactions will enable us to dissect this interconnected network to the “network motifs” of regulation as suggested by Shen-Orr et al. (6). Nevertheless, the network created in this study can be of great help in the analysis of a transcriptomic response by displaying the data on the constructed network. Moreover, combining the knowledge in this network and imposing it on the recently developed metabolic network of *L. plantarum* (34) will help to increase our understanding of global gene expression and metabolic adaptation of *L. plantarum* to changes in its dynamic environment.

References

1. **Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A. et al. (2003)** Computational discovery of gene modules and regulatory networks. *Nat Biotechnol, 21, 1337-1342.*
2. **Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. and Palsson, B.O. (2004)** Integrating high-throughput and computational data elucidates bacterial networks. *Nature, 429, 92-96.*
3. **Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. (2002)** Combining location and expression data for principled discovery of genetic regulatory network models. *Pac Symp Biocomput, 437-449.*
4. **Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I. et al. (2002)** Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science, 298, 799-804.*
5. **Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004)** Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature, 431, 308-312.*
6. **Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002)** Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet, 31, 64-68.*
7. **Gutierrez-Rios, R.M., Rosenblueth, D.A., Loza, J.A., Huerta, A.M., Glasner, J.D., Blattner, F.R. and Collado-Vides, J. (2003)** Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome Res, 13, 2435-2443.*
8. **Herrgard, M.J., Covert, M.W. and Palsson, B.O. (2004)** Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol, 15, 70-77.*
9. **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 U S A, 95, 14863-14868.*
10. **Thompson, W., Rouchka, E.C. and Lawrence, C.E. (2003)** Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res, 31, 3580-3585.*
11. **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.*
12. **Bussemaker, H.J., Li, H. and Siggia, E.D. (2001)** Regulatory element detection using correlation with expression. *Nat Genet, 27, 167-171.*
13. **Conlon, E.M., Liu, X.S., Lieb, J.D. and Liu, J.S. (2003)** Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A, 100, 3339-3344.*
14. **Keles, S., van der Laan, M. and Eisen, M.B. (2002)** Identification of regulatory elements using a feature selection method. *Bioinformatics, 18, 1167-1175.*
15. **Wels, M., Francke, C., Kerkhoven, R., Kleerebezem, M. and Siezen, R.J. (2006)** Predicting cis-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res, 34, 1947-1958.*
16. **Bailey, T.L. and Gribskov, M. (1998)** Combining evidence using p-values: application to sequence homology searches. *Bioinformatics, 14, 48-54.*
17. **Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W. et al. (2003)** Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A, 100, 1990-1995.*
18. **Casjens, S. (2003)** Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol, 49, 277-300.*
19. **Risoen, P.A., Johnsborg, O., Diep, D.B., Hamoen, L., Venema, G. and Nes, I.F. (2001)** Regulation of bacteriocin production in *Lactobacillus plantarum* depends on a conserved promoter arrangement with consensus binding sequence. *Mol Genet Genomics, 265, 198-206.*
20. **Martinez-Antonio, A. and Collado-Vides, J. (2003)** Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol, 6, 482-489.*
21. **Chauvaux, S., Paulsen, I.T. and Saier, M.H., Jr. (1998)** CcpB, a novel transcription factor implicated in catabolite repression in *Bacillus subtilis*. *J Bacteriol, 180, 491-497.*
22. **Francke, C., Kerkhoven, R., Wels, M. and Siezen, R.J. (2007)** Local and global transcriptional regulation and the potential for cross-talk, exemplified by the LacI transcription factor family in *Lactobacillus plantarum* WCFS1. *submitted for publication*.
23. **Derre, I., Rapoport, G., Devine, K., Rose, M. and Msadek, T. (1999)** ClpE, a novel type of HSP100

ATPase, is part of the CtsR heat shock regulon of *Bacillus subtilis*. *Mol Microbiol, 32, 581-593.*

24. **Kruger, E. and Hecker, M. (1998)** The first gene of the *Bacillus subtilis clpC* operon, *ctsR*, encodes a negative regulator of its own operon and other class III heat shock genes. *J Bacteriol, 180, 6681-6688.*
25. **Chastanet, A., Fert, J. and Msadek, T. (2003)** Comparative genomics reveal novel heat shock regulatory mechanisms in *Staphylococcus aureus* and other Gram-positive bacteria. *Mol Microbiol, 47, 1061-1073.*
26. **Chastanet, A., Prudhomme, M., Claverys, J.P. and Msadek, T. (2001)** Regulation of *Streptococcus pneumoniae clp* genes and their role in competence development and stress survival. *J Bacteriol, 183, 7295-7307.*
27. **Varmanen, P., Ingmer, H. and Vogensen, F.K. (2000)** ctsR of *Lactococcus lactis* encodes a negative regulator of *clp* gene expression. *Microbiology, 146 (Pt 6), 1447-1455.*
28. **Kelley, W.L. (2006)** Lex marks the CtsR heat shock regulon of *Bacillus subtilis*. *Mol Microbiol, 62, 581-593.*
29. **Kruger, E. and Hecker, M. (2003)** Identifying global regulators in transcriptional regulatory networks in bacteria. *J Bacteriol, 6, 482-489.*
30. **Kelley, W.L. (2006)** Lex marks the spot: the virulent side of SOS and a closer look at the LexA regulon. *Mol Microbiol, 62, 1228-1238.*
31. **Au, N., Kuester-Schoeck, E., Mandava, V., Bothwell, L.E., Canny, S.P., Chachu, K., Colavito, S.A., Fuller, S.N., Groban, E.S., Hensley, L.A. et al. (2005)** Genetic composition of the *Bacillus subtilis* SOS system. *J Bacteriol, 187, 7655-7666.*
32. **Fernandez De Henestrosa, A.R., Ogi, T., Aoyagi, S., Chafin, D., Hayes, J.J., Ohmori, H. and Woodgate, R. (2000)** Identification of additional genes belonging to the LexA regulon in *Escherichia coli*. *Mol Microbiol, 35, 1560-1572.*
33. **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, D75-77.*
34. **Teusink, B., Wiersma, A., Molenaar, D., Francke, C., de Vos, W.M., Siezen, R.J. and Smid, E.J. (2006)** Analysis of growth of *L. plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. *J Biol Chem, 281, 40041-40048.*

CHAPTER 8

General discussion and Future Perspectives

Acknowledgements
Thanks to Maria Marco and Maaike de Vries for sharing and discussing their microarray data.

Scope of this thesis:

The focus of this thesis was to dissect the regulatory network in *L. plantarum* using different bioinformatics tools. Although the bioinformatics approaches dominated, in many cases the generated hypotheses were validated using evidence gathered from functional genomics experiments. This chapter summarizes the main results and conclusions drawn from these results. In addition, it offers an outlook on the possible future research directions of this field.

Summary of the main results:

We have analyzed the regulatory network of *L. plantarum* using four different approaches, one based on the annotation of the regulatory proteins encoded on the genome, while the three other methods predict conserved *cis*-regulatory elements in the sequences upstream of genes or Transcriptional Units (TUs).

Identification and analysis of the regulatory proteins is the first step in understanding the regulatory network underlying the adaptive mechanisms of a unicellular organism (Chapter 2). The putative regulatory proteins of *L. plantarum* were subjected to advanced annotation and divided into (sub)-families based on the data of different genome databases (COG, Pfam and ERGO). Putative locally acting regulons were predicted using genome-context conservation and complete genome hybridization (CGH) data. The predicted regulons were grouped into classes based on biochemical function. Regulatory proteins within the same (sub)-family often appeared to regulate regulons containing genes with comparable molecular functions. Comparison with other lactic acid bacteria (LAB) genomes revealed that *L. plantarum* has the highest relative fraction of its predicted proteome assigned to regulatory proteins. The highest number of differences in LAB between the relative sizes of the regulatory protein families was observed for LacI, LysR and MarR. These regulator families can generally be correlated to adaptation to different or changing environments, i.e. regulation of energy metabolism and cell-envelope processes in relation to available nutrients (LacI and LysR), and regulation of protective mechanisms in response to harmful chemicals (MarR). The high number of regulatory proteins within these three regulatory families is in line with the flexibility and versatility of the species *L. plantarum*.

In the other chapters, regulons were predicted by employing three different methods to identify conserved regulatory elements in the upstream regions of the *L. plantarum* genes (Figure 1); i) large-scale global motif searching in the upstream regions of all genes, ii) phylogenetic footprinting on basis of conserved Transcriptional Units (TUs) and iii) clustering and iteration of TUs with transcriptomics-based correlated expression. Each method approached the ambition to predict the gene-regulation network of *L. plantarum* from a different angle, and employed a different source of information to initiate the regulatory element search; i) the annotated genome sequence of *L. plantarum* WCFS1, ii) the annotated genome sequences of different sets of related species, and iii) multiple transcriptome datasets from a variety of experiments performed with *L. plantarum* WCFS1.

In the first method, the upstream regions of all genes were searched for the presence of the best-conserved motifs in the genome (Chapter 3). Of the ten most commonly found motifs, nine showed to be similar to intergenic elements described in other bacteria. Seven of these known motifs, as well as the only new identified motif, are involved in processes related to transcription or translation. The remaining three motifs appear to be part of

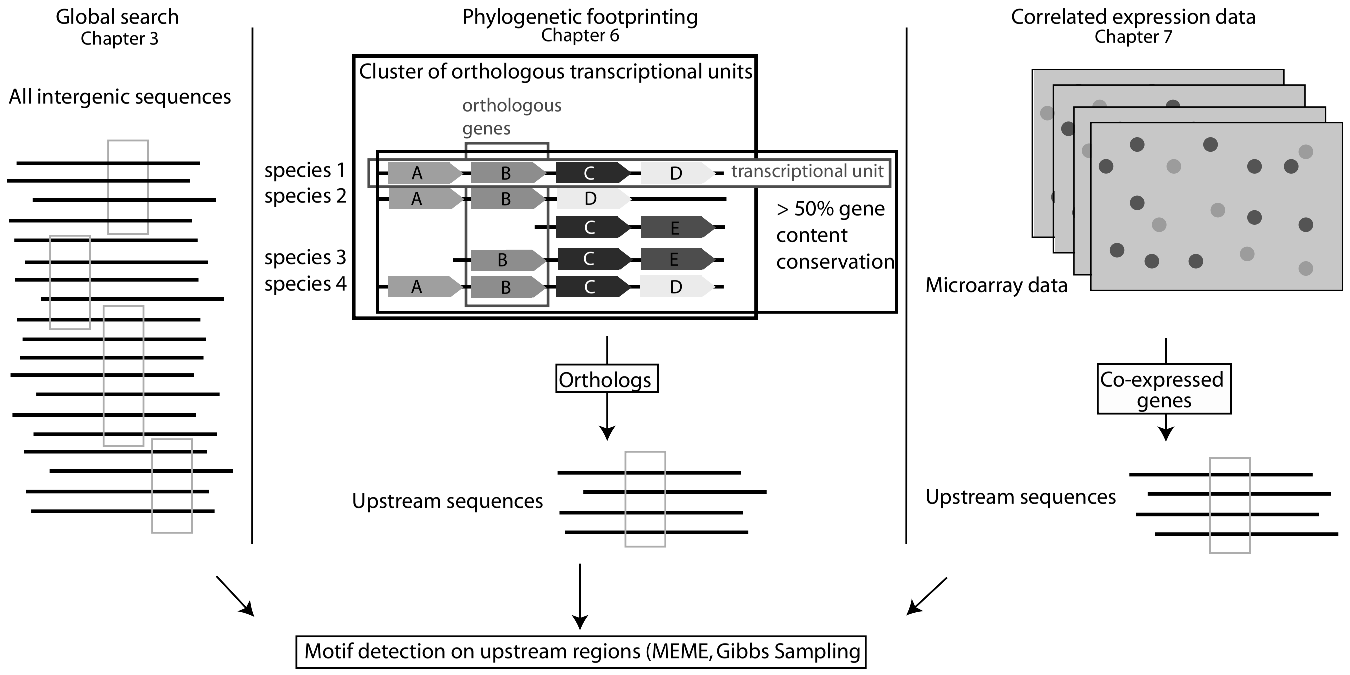


Figure 1 (a color version of the figure can be found at page 184): different procedures used to predict regulatory elements in *L. plantarum*.

the inverted repeats flanking transposons of the previously identified types ISP1 and ISP2 (1). Co-occurrence of two motifs in the same upstream regions appeared to result from separate identification of different parts of the T-box regulatory element (analyzed in detail in Chapter 5), and from different representations of the sigmaA (SigA) promoter. These promoter elements were used to predict the full complement of SigA promoters on the genome of *L. plantarum*. By using the observed spatial preference of the promoter in relation to the start codon of the downstream-located gene, promoters could be predicted with a significance level of 85%. Applying this significance cut-off allowed the identification of SigA-dependent promoters in the upstream regions of 45% of the predicted transcriptional units. An explanation for the low specificity in the identification of the SigA-promoter element is the possible overlap of the promoter sequences with additional regulatory elements that activate or repress binding of the RNA polymerase to the DNA.

The second method to predict *cis*-acting elements in *L. plantarum* was by comparative analysis of the upstream regions of conserved genes and predicted TUs in different bacterial genomes (Chapter 6). TUs were predicted for two species sets, with different evolutionary distances to *L. plantarum*. TUs were designated to constitute a “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 could be demonstrated that phylogenetic footprinting using different species sets can identify and distinguish between general regulatory motifs and LAB-specific regulatory motifs.

In the third method, transcriptome data was used as the basis of motif prediction (Chapter 7). A combination of transcriptome data from over 70 different microarray experiments was used to elucidate relations between the activity of certain transcription factors (TFs) and the genes they control, thereby establishing the regulatory network of a cell. The TUs that displayed correlated expression regulation were used to identify putative *cis*-regulatory elements by searching the upstream regions of the TUs for conserved motifs. Predicted motifs were validated by searching for additional motifs in the upstream regions of other TUs with correlated expression regulation as compared to the three initial TUs. The upstream regions of these TUs were used to refine the putative *cis*-acting elements. In total, *cis*-acting elements were identified for 41 different regulons consisting of at least 4 highly correlated TUs (correlation >0.7). It was found that the regulatory network of *L. plantarum* is highly interconnected and individual TUs were found to be a member of up to 6 different regulons. In addition to these large-scale genomics approaches to identify *cis*-regulatory elements, two of the motifs found in these large-scale analyses were studied in more detail

(Chapter 4 and 5). Both these elements are transcribed and probably display their function as an RNA molecule.

Inspection of the large (>700 nt) intergenic regions of the *L. plantarum* WCFS1 genome sequence revealed the existence of several highly conserved, genetically linked motifs (8 – 34 nt) that form multiple mosaic *L. plantarum* supermotifs (LPSM) (Chapter 4). A Hidden Markov Model (HMM) was used to refine LPSM detection, revealing 24 LPSM occurrences with a length varying from ~800 to 1000 nt. Secondary structure analyses predicted conserved cruciform-like structures for these LPSM. The LPSMs appear to be unique for *L. plantarum* and were found to be highly conserved among different *L. plantarum* strains by DNA-microarray analyses and locus-specific PCR amplification and sequencing. Northern blot analysis and meta-analysis of transcriptome data sets indicated transcription of these LPSMs. In addition, transcriptome data suggested regulation of the expression in experiments comparing *L. plantarum* WCFS1 wild-type strain with an isogenic strain over-expressing the endogenous thioredoxin reductase-encoding *trxB1*. Intriguingly, similar regulatory effects were observed in other studies comparing *L. plantarum* WCFS1 and other strains over-expressing genes *in trans*. Although the function of these LPSMs remains unknown, they could play a role as regulatory RNAs acting as a defense system against the presence of genes of which the expression could be detrimental to the cell.

Phylogenetic footprinting, motif searching and RNA structure prediction procedures were employed to analyze the presence of T-box elements and their specifier codons (Chapter 5). This method was applied to all currently sequenced prokaryotic genomes. Despite their supposed ancient occurrence in bacteria, the taxonomic distribution of T-boxes appeared remarkably restricted. T-box elements are abundantly present in Firmicutes species, with an average of sixteen representatives per genome, whereas only ten species outside the phyla of Firmicutes and Actinobacteria were found to contain T-boxes. Our findings confirmed the current view that the vast majority of the T-box regulated genes (>92%) is directly related to charging uncharged tRNA (i.e. via tRNA ligation, amino acid biosynthesis and amino acid transport). The specifier codon of the various T-box elements was used to improve the functional annotation of approximately 125 genes, including many genes that are notoriously difficult to annotate on basis of sequence similarity like amino acid transporters. A detailed analysis of especially this group of T-box elements indicated that the T-box should behave as an independently evolving functional module and can easily switch amino acid specificity. We hypothesize that the ancestral T-box was linked to one of the genes encoding a tRNA ligase of a branched-chain amino acid, probably Ile.

Future perspectives

The availability of complete genome sequences of organisms has changed the way we look at the functioning of an organism. While previous studies have focused on the properties of a single gene or protein, genomics has opened the possibility to study the full complement of genes on the genome. Furthermore, functional genomic techniques, such as transcriptomics, allow to study interactions between genes and identify novel links between seemingly unrelated genes. Next to the implementation of the genome sequence in the development of post-genomics methods, the availability of a complete genome sequence has lead to attempts to create different *in silico* models of the cell (2). The construction of metabolic as well as transcriptional networks has lead to an integrative view on genomes by assessing the function of a gene to the context it fulfills in the cell. Recently a genome-scale metabolic model was created for *L. plantarum* (3). This model links gene products to metabolic reactions that together lead to a mathematical description that is used to analyze and predict how *L. plantarum* will grow when challenged with different constraints (limitations). The development of this metabolic model is the first step in creating an *in silico* cell, a mathematical description of an organism that can mimic the reaction to a change in the environment. Integration of the data on the regulatory network of *L. plantarum* generated in this thesis will be the second step in the expansion of the *in silico* cell model. Although only a few attempts have been made to integrate the properties of metabolic models with knowledge regarding the regulation of gene expression were met with limited success, these first results show that metabolism and regulation are highly intertwined (4).

Motif prediction

Four different methods for predicting regulons were evaluated in order to predict a complete regulatory network for *L. plantarum*. Although one might expect to observe a significant overlap between the different predictions, our results are better characterized by the vast amount of differences we observed. In some of the well-described examples, this limited amount of overlap can be explained on basis of biological knowledge. As an example, the HrcA-binding site predicted from the phylogenetic footprinting (Chapter 6) was found to control the expression of only two transcriptional units. Although these two TUs show a clearly correlated expression, they were not found using the microarray data based analysis (Chapter 7), because this microarray set required at least three coregulated TUs in order to start motif prediction. On the other hand, regulons identified on basis of correlated expression data, like the SOS-regulon, were not identified with the phylogenetic footprinting. In this case, it was shown that the SOS-regulons of different species share a lot of genes of comparable function, but only a few true orthologs. Because phylogenetic footprinting assumes co-regulation of orthologous genes (or TUs), regulons lacking shared orthologs between species will be missed. Therefore, use of methods like phylogenetic footprinting and expression correlation data on their own will still lead to a high number of false-positive *cis*-acting elements.

On basis of the research presented in this thesis one could conclude that there is no generally best-suited method for predicting regulons in bacteria. The best possible regulon prediction is probably reached by combining the information gathered from different resources that include phylogeny of the regulatory protein, phylogeny of the TUs, large-scale genome sequence-motif mining, microarray and literature data. We recently showed that combining knowledge of genome context, phylogeny and microarray data could result in a high-quality regulon prediction of the different members of the LacI family within *Firmicutes* (5). This study employed the LacI family of transcriptional regulators as a test case to show that using the resources described above, discrimination between the *cis*-acting elements can be achieved, thereby dividing the general *cre* element into different individual members, and allowing specific regulon prediction for all the LacI family members. This procedure is applicable in a broader sense and has been shown to be successful for other families of regulatory proteins, like two-component systems (M. de Been, manuscript in preparation).

Microarray data

Although the power of the applied methods largely depended on the availability of transcriptomics data, a large fraction of regulatory elements could not be validated (Chapter 5) or integrated in the initial analysis (Chapter 7) due to the lack of sufficient high-quality data. A large fraction of the genes (685 genes, about 22% of the total genome) were either not spotted on the array (216 genes) or excluded for expression correlation analysis due to inconsistencies in the expression patterns of the different probes of a single gene. However we observed that for many of these probes, both the measured signal intensities as well as the observed change in expression ratio were very low, causing the noise to be a significant fraction of the total signal.

For 1591 out of 2393 genes suitable for expression correlation analysis, the expression ratio (^{2}log ratio, M) in the performed experiments did not change more than within the cut-off boundaries set at -1 to +1, suggesting that these genes are either never expressed, or constitutively expressed at the same level in all the conditions tested. Although the microarray set used is of a reasonable size (72 experiments), all data was gathered on basis of *L. plantarum* grown under laboratory conditions in either batch or continuous cultures but using a limited number of different media. To obtain further insight in the expression of these “unaffected genes”, transcriptome experiments of *L. plantarum* in other environments are likely to increase the number of significantly expressed/regulated genes and, thereby, improve the resolution of the analyses performed. In recent studies by Marco et al. (in preparation) and de Vries et al. (6), *L. plantarum* was introduced into the GI tract of mammals (mouse and human, respectively), an ecosystem that is very different from any other (laboratory) condition tested to date. Analysis of these two data sets revealed that 147 and 348 of the above mentioned 1591 “unaffected genes” were differentially up-regulated in the mice and man data sets, respectively, as compared to any of the studied laboratory experiments. Comparison of the two data sets revealed that 31 of these genes were up-regulated in both mouse and man gastrointestinal tracts (Table1). Although the *L. plantarum* expression data was obtained from bacterial cells of different origin (mouse *vs* human) and sampled from different parts of the intestine (ileum and colon in the human studies *vs* caecum in the mouse studies), there seems to be a common genetic response to the passage through the mammalian GI tract. Although the number of shared genes is comparable to what can statistically be expected to be shared between to unrelated set of random genes, many genes can plausibly be related to GI-tract adaptation. Six genes were found to be related to transport and binding of different substrates: 3 for carbohydrates (including one PTS system), one for amino acids and two with an unknown substrate. In addition, genes with functions related to “Cell envelope” (6 genes), “Energy metabolism” (4 genes), and “Hypothetical proteins” (8 genes) were included in this group of 31 genes. Cell-envelope proteins can be expected to be up-regulated in response to intestinal conditions (7-9) and potentially play a role in attachment to diet- or host-derived components of the GI-tract. The presence of energy metabolism genes (see Table 1) could indicate that *L. plantarum* has changed its metabolism compared to the laboratory conditions.

Next to the identification of *in vivo* up-regulated genes, the mouse intestine -derived data sets were also suitable for identification of genes that display clear *in vivo* down-regulation. Similar analyses were not possible for the human intestine-derived transcriptome datasets of *L. plantarum,* which is due to the considerably lower transcription detection limit and other technical limitations associated with these experiments. In total, 253 out of the 1591 genes were significantly down-regulated in the mouse GI-tract as compared to laboratory conditions. Most of these genes are of unknown function (68 classified as hypothetical), followed by genes involved in transport and binding (38 genes). In addition, many of the down-regulated genes were found to be involved in general maintenance processes of the cell like protein synthesis (31 genes), energy metabolism (12 genes) and protein fate (12 genes). An interesting observation was that 19 out of 31 down-regulated genes classified as "Protein synthesis" were involved in tRNA aminoacylation or base modification, suggesting that protein synthesis is down-modulated in the GI-tract. This observation is supported by the observation that also many protease- and peptidase-encoding genes (classified as 'Protein fate') appeared to be down-regulated *in vivo*. Another striking and possibly somewhat unexpected observation is the down-regulation of a relatively high number of genes encoding cell-envelope proteins (22 genes in total). The change in expression of these genes, together with the observation that 15 genes related to "Cell-envelope" are upregulated within the mouse intestine-derived dataset, indicates that a shift to a completely different environment dramatically changes the makeup of the cell wall. In total, the two *in vivo* data sets included in the analysis presented here increased the number of differentially expressed genes with 654, which is 45% of the 1591 previously non-differentially expressed genes and shows that the value of expression correlation analysis increases drastically with the addition of more (diverse) data sets that go beyond the standard laboratory conditions.

Since the microarray platform used in these studies was a two-color microarray, the main focus in these studies was on genes that show a change in expression ratio (M-value) throughout different experiments. Although we observed many genes with a changed expression ratio between the two compared conditions, the data is always limited by the comparison of the individual conditions in one microarray. As an example, in all of the experiments, medium composition and fermentation conditions were identical between the two conditions, where medium composition and fermentation conditions did differ between all performed experiments. These differences cannot be analyzed when only the M-value between the conditions within one experiment are used for expression correlation analysis. For effective use of all variance in the experiments, also between the different experiments, the intensities of the different spots per condition would have to be taken into account. Although these data was accessible to us, it proved hard to distillate biological meaningful information out of these data, mainly due to a high variance between the measured signal on the different microarray slides and lack of suited signal normalization algorithms. Statistical analysis of these microarray data showed that the major differences between the individual spot signals were not caused by biological variability of the conditions, but by technical issues, for example the individual that performed the specific microarray (Figure 2), technical noise that is reduced in the M-value by proper data normalization.

The second cause of a high number of genes being excluded from the correlation and validation analyses was caused by a lack of or discrepancies between gene-specific probes of a particular gene. Stringent criteria were set to probe selection procedures to reach a high reliability of the measured gene expression

Table 1: Genes induced in both GI-tract experiments compared to all tested lab conditions, sorted by functional classification

ORF	gene name	product	main class	sub class
lp_0209		adherence protein, C-terminal fragment	Cell envelope	Cell surface proteins: other
lp_1449		extracellular protein	Cell envelope	Cell surface proteins: other
lp_2978		extracellular protein	Cell envelope	Cell surface proteins: other
lp_3067		extracellular protein	Cell envelope	Cell surface proteins: other
lp_3074		cell surface protein precursor	Cell envelope	Cell surface proteins: LPxTG anchor
lp_3452		extracellular protein	Cell envelope	Cell surface proteins: other
lp_1701		nucleotide-binding protein, universal stress protein UspA family	cellular processes	adaptation and atypical conditions: other
lp_3641	*dexB*	glucan 1,6-alpha-glucosidase	Central intermediary metabolism	Biosynthesis and degradation of polysaccharides
lp_0506	*sdhA*	L-serine dehydratase, alpha subunit	Energy metabolism	Amino acids and amines
lp_1471	*nifU*	NifU-like protein	Energy metabolism	Electron transport
lp_3480	*galT*	UTP--hexose-1-phosphate uridylyltransferase	Energy metabolism	Sugars
lp_3491		fumarate reductase, flavoprotein subunit precursor	Energy metabolism	Electron transport
lp_0168	*dak1B*	glycerone kinase	Fatty acid and phospholipid metabolism	Glycerolipid metabolism
lp_0611		unknown	Hypothetical proteins	Conserved: other
lp_1168		unknown	Hypothetical proteins	Not conserved: other
lp_2766		unknown	Hypothetical proteins	Conserved: other
lp_2767		acetyltransferase (putative)	Hypothetical proteins	Conserved: putative function
lp_2792		oxidoreductase	Hypothetical proteins	Conserved: putative function
lp_3301		integral membrane protein (putative)	Hypothetical proteins	Conserved: membrane proteins
lp_3427		metal-dependent hydrolase (putative), C-terminal fragment	Hypothetical proteins	Conserved: putative function
lp_3632		unknown	Hypothetical proteins	Conserved: other
lp_1904	*pepT*	tripeptidase	Protein fate	Degradation of proteins, peptides, and glycopeptides
lp_0501	*serS1*	serine--tRNA ligase	Protein synthesis	tRNA aminoacylation
lp_2963		transcription regulator (putative)	Regulatory functions	TetR/AcrR-family regulators
lp_3649		transcription regulator	Regulatory functions	LacI-family regulators

ORF	gene name	product	main class	sub class
lp_0178	*malA*	maltose/maltodextrin ABC transporter subunit (putative)	Transport and binding proteins	Carbohydrates, organic alcohols and acids
lp_1263	*oppC*	oligopeptide ABC transporter, permease protein	Transport and binding proteins	Amino acids, peptides and amines
lp_1393		ABC transporter, ATP-binding and permease protein	Transport and binding proteins	Unknown substrate
lp_2739		ABC transporter, ATP-binding protein	Transport and binding proteins	Unknown substrate
lp_3008	*pts23A*	cellobiose PTS, EIIA	Transport and binding proteins	PTS systems
lp_3643		sugar ABC transporter, permease protein	Transport and binding proteins	Carbohydrates, organic alcohols and acids

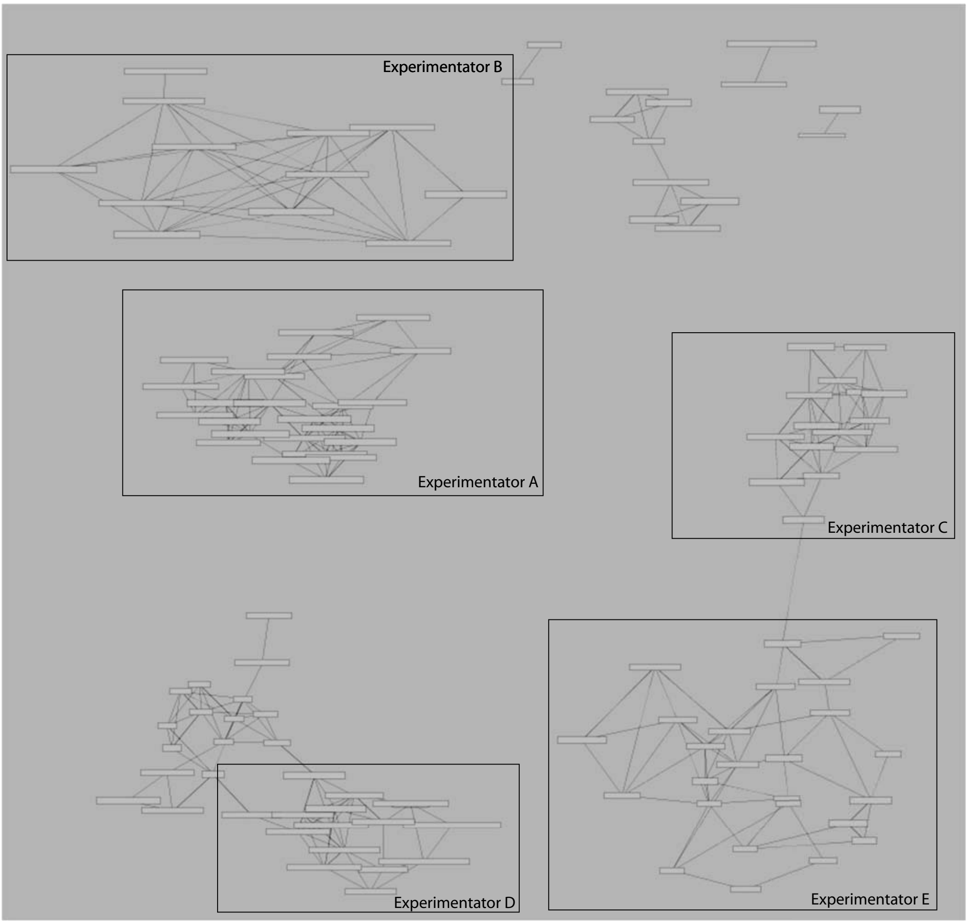


Figure 2: Clustering analysis of microarray channel data.
White boxes represent individual channel intensity measurements of different microarray experiments clustered by the fraction of shared highly expressed genes. Black boxes indicate experiments performed by one individual (A, B, C, D or E).

ratios (for details see Appendix). Although the different probes were designed to give a comparable and selective gene-specific signal with a low-noise component, the correlation analysis revealed that in several cases these probes do not all accurately report the correlated expression of individual genes (<0.4) throughout the performed experiments. This low correlation caused 457 of the genes to be unsuitable for the expression-correlation analysis. Recent developments in microarray expression platforms have allowed the synthesis of slides with 244,000 probes spotted on one microarray slide. This new technology was used to represent the entire genome sequence of *L. plantarum* on two Agilent microarray slides representing the two strands of the genome. The design of this microarray was based on a probe-tiling strategy (partially overlapping probes), thereby increasing resolution to detect potential small deletions (10 – 20 nt) in the genome. These microarrays can be used for analysis of the transcriptome at a much higher level of resolution. Probes that give a reliable signal compared to other probes can be candidates to use in an improved, three or four probe per gene microarray design.

In addition this full-coverage array design has more advantages. As the microarray covers the complete genome, potential regulatory RNA molecules can be detected. Several RNA molecules predicted to be present in the different chapters of this thesis can be validated with experimental data from these tiling arrays. Predicted T-box positions (Figure 3C) and LPSM locations (Figure 3B) were shown to be part of the transcriptome of *L. plantarum*. Predictions regarding the promoter locations of genes could also be validated; this could even be achieved for some promoters that were regarded false-positive on basis of their distance to the translation start of the first downstream-located gene (Figure 3A). Further analysis of the predicted promoter regions will be performed in the future to validate the predicted promoter regions by evaluating if the predicted promoter truly points at a transcription start point. Transcription start sites can be identified by observing an increased signal intensity starting directly downstream of the predicted promoter. In addition, the use of high density tiling microarrays will allow for the identification of regulatory RNAs.

Final Remarks

Although the proposed regulatory network will prove valuable in understanding the adaptation response of *L. plantarum*, additional work will need to be performed to increase the value of the network. Probably the biggest drawback of the network is the lack in experimental data describing the interactions of the transcription factor (TF) with their *cis*-acting elements. TF-binding data from for example ChIP-chip experiments will lead to the prediction of a regulatory network with a higher resolution. Although we observe a degree of connectivity that is comparable to earlier observations in other bacteria, detailed information on TF-*cis*-regulatory element interactions will enable us to dissect this interconnected network to the “network motifs” of regulation as suggested by Shen-Orr et al.(10). Nevertheless, the network created in this study can be of great help in the analysis of a transcriptomic response by displaying the data on the constructed network. Moreover, combining the knowledge in this network and imposing it on the recently developed metabolic network of *L. plantarum* (3) will help to increase our understanding of the global gene expression and metabolic adaptation of *L. plantarum* to changes in its dynamic environment.

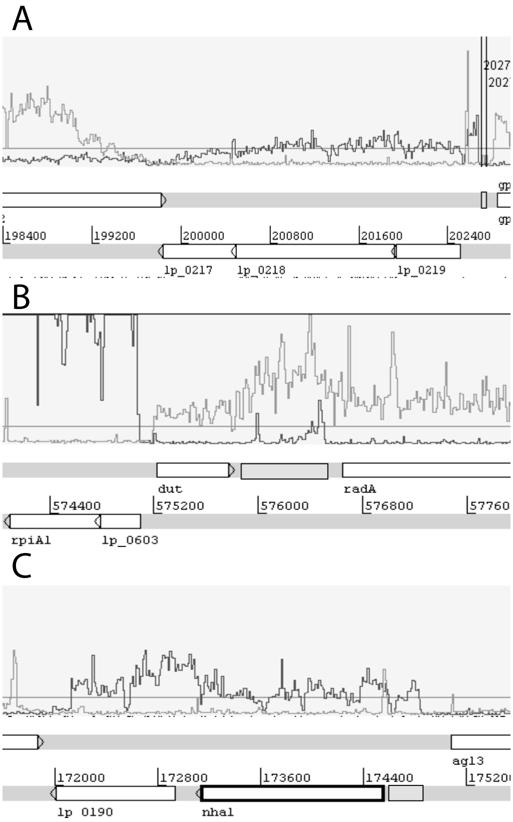


Figure 3 (a color version of the figure can be found at page 185): Different examples of the use of tiling microarray data to validate transcriptome predictions.

Expression levels are shown for the + (green) and – (red) strand. A) shows the start of transcription of a TU approximately 200 nt upstream of the first gene of the TU (*lp_0219*), near the predicted promoter (202705 – 202746, indicated in yellow) and continues to the end of the TU (*lp_0217*). B) shows the transcription of one of the LPSMs (indicated in yellow) as being part of a transcript that starts upstream of dut and continues expression into radA. C) shows the transcription of a T-box (indicated in yellow) in the upstream region of *lp_0191* (annotated as a natrium antiporter (Nha1) in the original *L. plantarum* annotation).

References

1. **Kleerebezem, M., Boekhorst, J., van Kranenburg, R., Molenaar, D., Kuipers, O.P., Leer, R., Tarchini, R., Peters, S.A., Sandbrink, H.M., Fiers, M.W. et al. (2003)** Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A, 100, 1990-1995.*
2. **Stelling, J. (2004)** Mathematical models in microbial systems biology. *Curr Opin Microbiol, 7, 513-518.*
3. **Teusink, B., Wiersma, A., Molenaar, D., Francke, C., de Vos, W.M., Siezen, R.J. and Smid, E.J. (2006)** Analysis of growth of *Lactobacillus plantarum* WCFS1 on a complex medium using a genome-scale metabolic model. *J Biol Chem, 281, 40041-40048.*
4. **Kharchenko, P., Church, G.M. and Vitkup, D. (2005)** Expression dynamics of a cellular metabolic network. *Mol Syst Biol, 1, 2005 0016.*
5. **Francke, C., Kerkhoven, R., Wels, M. and Siezen, R.J. (2007)** Local and global transcriptional regulation and the potential for cross-talk, exemplified by the LacI transcription factor family in *Lactobacillus plantarum* WCFS1. *submitted for publication.*
6. **de Vries, M.C., Marco, M.L., Kleerebezem, M., Mangell, P., Ahrne, S., Molenaar, D., de Vos, W.M. and Vaughan, E.E. (2007)** Transcript profiling reveals global gene expression of a probiotic *Lactobacillus plantarum* in the human intestinal tract. *submitted for publication.*
7. **Bron, P.A., Grangette, C., Mercenier, A., de Vos, W.M. and Kleerebezem, M. (2004)** Identification of *Lactobacillus plantarum* genes that are induced in the gastrointestinal tract of mice. *J Bacteriol, 186, 5721-5729.*
8. **Denou, E., Berger, B., Barretto, C., Panoff, J.M., Arigoni, F. and Brussow, H. (2007)** Gene Expression of a Commensal *Lactobacillus johnsonii* Strain NCC533 during In Vitro Growth and in the Murine Gut. *J Bacteriol. 189(22):8109-8119*
9. **Sonnenburg, J.L., Xu, J., Leip, D.D., Chen, C.H., Westover, B.P., Weatherford, J., Buhler, J.D. and Gordon, J.I. (2005)** Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science, 307, 1955-1959.*
10. **Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002)** Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet, 31, 64-68.*

APPENDIX

Correlex, a web-based tool for studying gene expression correlation in *L. plantarum*

Acknowledgements

Thanks to Lex Overmars (Student HBO Bioinformatics, HAN) for the original design and implementation of **Correlex**. Thanks to Douwe Molenaar for insightful discussion, probe design and supplying MySQL dumps of the BASE database.

Scope

This appendix describes **Correlex**, a database and web-based tool for searching and visualizing correlated expression of genes over different microarray data sets in *L. plantarum*.

Expression correlation

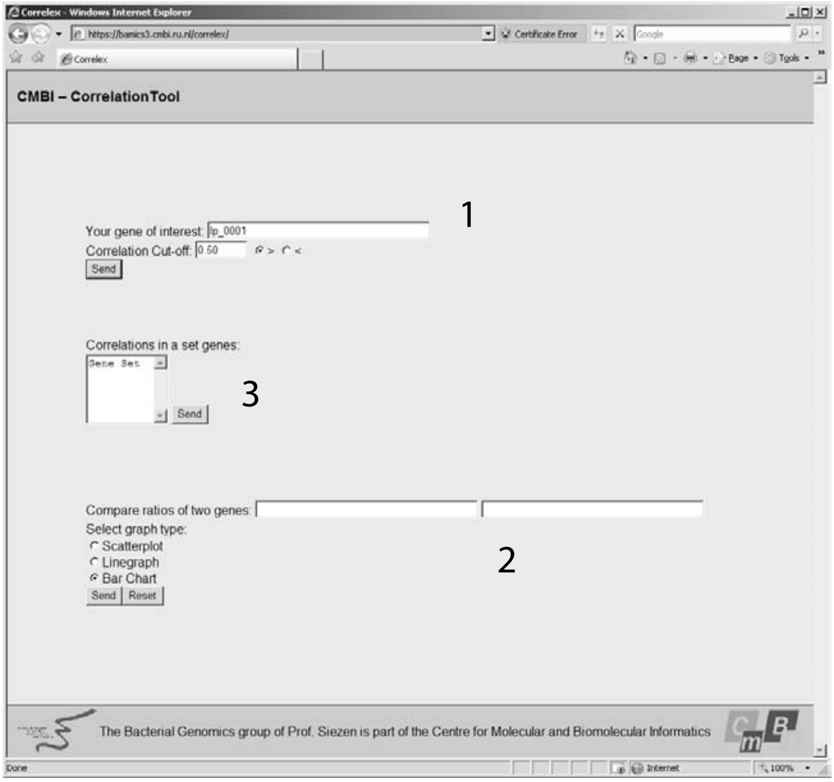
Transcriptomics data gathered from microarrays is useful for studying the genetic response of an organism under a specific condition, i.e. how gene expression changes in response to external stimuli. In addition, combining the data from different microarray experiments can be used to create a gene-specific expression profile over multiple experiments. Comparing these profiles by calculation of the expression correlation can enable the prediction and/or validation of co-regulation; genes that have a correlated expression are likely to be co-regulated. **Correlex** is a database that uses the expression data of different transcriptomics experiments of *L. plantarum* to calculate the expression correlation of genes and displays this correlation-data in comprehensive manner through a user-friendly (web-)interface. Researchers can use these expression correlations to identify co-regulated genes. This information can help them in their research, for example in improving the annotation of a gene of interest by linking its expression profile to that of one or more gene(s) of known function (GEO reference GPL4318).

Microarray data

The availability of the complete genome of *L. plantarum* has led to the construction of microarrays that represent all (or almost all) the genes encoded on the genome. During the last 5 years, several different microarray platforms have been used, based on the technology available at that time. The most recent technological platform used for microarray experiments is based on gene-representing 60-mer oligonucleotide probes that are printed on a glass slide using ink jetprinter technology (Agilent: http://www.agilent.com). The current format of the microarray is such that approximately three different probes per gene are printed on the glass slide. These probes were designed to minimize cross-hybridization with other genes on the genome and to maximally equalize melting temperatures.

Calculation of the correlated expression

All transcriptomics data from the different experiments were first normalized by local fitting of an M-A plot using the implementation of the LOWESS algorithm in R (http://www.r-project.org). All expression correlations were calculated using the Pearson correlation coefficient (1). To get to gene-gene correlations, the mean expression ratio per gene was calculated on basis of the individual probe signals. The different oligonucleotide probes per gene were checked for consistent correlation before including the gene to the gene-gene analysis. Genes that lacked a consistent correlation between the different probes were removed from the data set. If all probe-probe correlations were above cutoff (0.4) the signals of all probes were used. In cases where only two out of three probes showed to have a correlated expression, the third probe was not regarded in further analyses and the gene signal was calculated on basis of the data from only two probes. If all probe-probe correlations were below 0.4, none of the probes were used. Although in these cases one probe could still represent the true expression pattern of the gene, it is impossible to decide which probe shows the correct expression pattern. All data was stored in a database (MySQL) and organized in a manner enabling frequent and easy updating without expert knowledge on the database system. At present, the BASE



CMBI – CorrelationTool

Your gene of interest: lp_0001

Correlation Cut-off: 0.50 ● > ○ <

Correlations in a set genes: Gene Set

Compare ratios of two genes:

Select graph type:

- ○ Scatterplot
- ○ Linegraph
- ● Bar Chart

The Bacterial Genomics group of Prof. Siezen is part of the Centre for Molecular and Biomolecular Informatics

Figure 1: Entry page of Correlex.
On this first page the user can select three different types of data analysis, based on a single gene (1), two genes (2), or a set of genes (3).

information system for storing microarray data (2) is used as a source for transcriptomics data. At the 1st of September 2007, **Correlex** contained correlated expression data of 2450 genes (including plasmid, rRNA and tRNA genes) of *L. plantarum* based on 72 different experiments.

Application

Correlex can be accessed from https://bamics3.cmbi.ru.nl/correlex/ and allows three types of analysis (Figure 1): i) get information of the best-correlating genes based on submitting a query gene of interest, ii) submitting a list of query genes and get insight on the correlation of all possible gene pairs in this list and iii) get correlated expression of two genes of interest and visualize the correlated expression in three different formats (Figure 2).

If the user submits the locus name of interest (*lp_number*) to **Correlex** (Figure 1, option 1) the next page (gene page) will provide the user with transcriptomics data on this gene (Figure 3). Clicking on the gene name will link to the up-to-date gene annotation in LacPlantDB (see appendix of (3)). The graph on the gene page shows the expression profiles of the individual probes of the gene. If an expression profile is represented by a dotted line, this probe was not regarded in

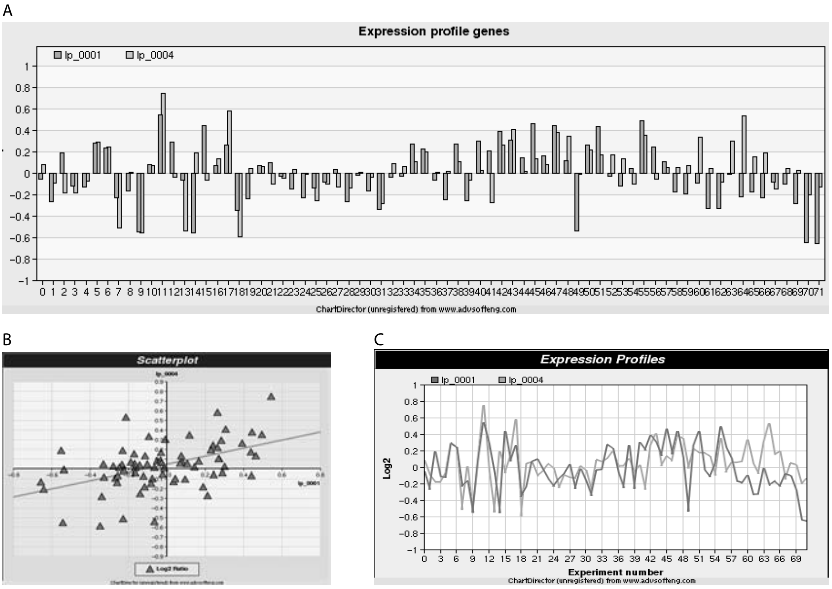
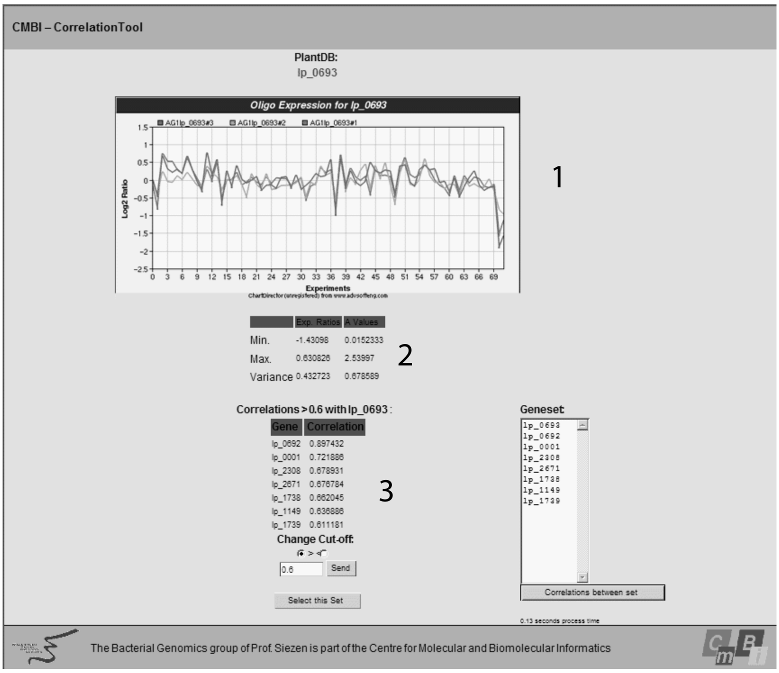


Figure 2: Different output formats for ratio comparisons of two genes: bar (A), scatter (B) and line (C) plot.



CMBI – CorrelationTool

PlantDB:
lp_0693

	Exp. Ratios	A Values
Min.	-1.43098	0.0152333
Max.	0.630826	2.53997
Variance	0.432723	0.678589

Correlations > 0.6 with lp_0693 :

Gene	Correlation
lp_0692	0.897432
lp_0001	0.721886
lp_2308	0.678931
lp_2671	0.676784
lp_1738	0.662045
lp_1149	0.636886
lp_1739	0.611181

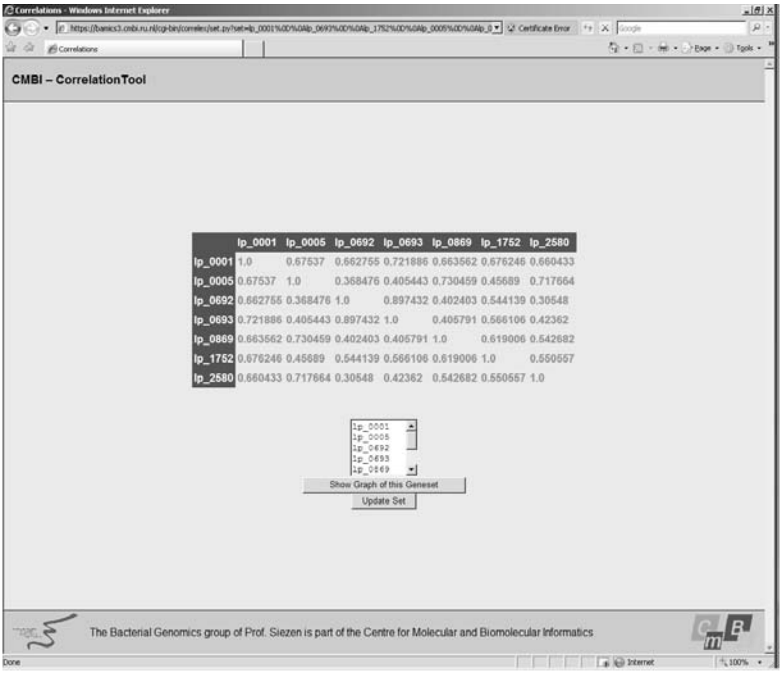
Change Cut-off:
● > ○ <
0.6

Geneset:
lp_0693
lp_0692
lp_0001
lp_2308
lp_2671
lp_1738
lp_1149
lp_1739

0.13 seconds process time

The Bacterial Genomics group of Prof. Siezen is part of the Centre for Molecular and Biomolecular Informatics

Figure 3: Gene page for *lp_0693*.
1) Expression profile of the three different gene-specific probes. 2) Statistical data regarding the gene of interest. 3) List of genes scoring above threshold with the gene of interest.


Correlations - Windows Internet Explorer
https://banics3.cmbi.ru.nl/cgi-bin/correlex/set.py?set=lp_0001%0D%0Alp_0693%0D%0Alp_1752%0D%0Alp_0005%0D%0Alp_0
Certificate Error
Google
Correlations
Page
Tools
Done
Internet
100%

CMBI – CorrelationTool

	lp_0001	lp_0005	lp_0692	lp_0693	lp_0869	lp_1752	lp_2580
lp_0001	1.0	0.67537	0.662755	0.721886	0.663562	0.676246	0.660433
lp_0005	0.67537	1.0	0.368476	0.405443	0.730459	0.45689	0.717664
lp_0692	0.662755	0.368476	1.0	0.897432	0.402403	0.544139	0.30548
lp_0693	0.721886	0.405443	0.897432	1.0	0.405791	0.566106	0.42362
lp_0869	0.663562	0.730459	0.402403	0.405791	1.0	0.619006	0.542682
lp_1752	0.676246	0.45689	0.544139	0.566106	0.619006	1.0	0.550557
lp_2580	0.660433	0.717664	0.30548	0.42362	0.542682	0.550557	1.0

The Bacterial Genomics group of Prof. Siezen is part of the Centre for Molecular and Biomolecular Informatics

Figure 4: Gene expression matrix showing the expression correlation of all possible different gene pairs.

further expression correlation calculation procedures due to a low correlation with other probes. Below the graph different statistical data for the gene of interest are shown (i.e. highest and lowest measured signals, level of variance observed in the expression profile). At the bottom, genes correlating above the selected threshold with the gene of interest are shown. On the right the user can select genes he/she wants to evaluate in the next steps. After submitting a list of genes, the next page shows these genes in an expression correlation matrix (Figure 4). This page with the expression correlation matrix can also be accessed directly from the start page by submitting a list of genes (Figure 1, option 2). Gene-gene correlations are colored according to the height of the correlation; red for correlations below 0.50 and green for correlations of 0.50 or above. Clicking on a correlation will open a gene-gene scatter plot (Figure 2B), clicking on a gene will open the corresponding gene page for that gene (Figure 3). Below the matrix, the user can find the option to add or remove genes from the matrix. Clicking on "Show graph" will show the expression profiles of all submitted genes in a line graph (shown for only two genes in Figure 2C).

Future

Recently, the **Correlex** system was also used to calculate and visualize expression correlation gathered from transcriptomics experiments of *Bacillus cereus* ATCC14579. Both systems will be frequently updated with new transcriptomics data. In addition, efforts will be made to make **Correlex** suited for expression data from previous and future microarray technologies and platforms.

References

1. **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 U S A, 95, 14863-14868.*
2. **Saal, L.H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, A. and Peterson, C. (2002)** BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol, 3, SOFTWARE0003.*
3. **Boekhorst, J. (2007),** Computational genomics of gram-positive bacteria, PhD thesis, Radboud University, Nijmegen.

Samenvatting voor iedereen (en in het Nederlands)

Als onderzoeker is het een van je taken om, aan iedereen die daarin geïnteresseerd is, te vertellen wat het belang van je werk is. Ik zal proberen om dat in dit hoofdstuk uit te leggen. Het onderzoek dat ik in de laatste 5 jaar heb uitgevoerd gaat over het bestuderen van melkzuurbacteriën, de werkpaarden van de zuivelindustrie. Waar veel mensen niet vaak bij stilstaan, is dat melkzuurbacteriën ons leven verrijken met heel veel lekkere en gezonde voedingsmiddelen zoals yoghurt en kaas, maar ook verschillende soorten worst, zuurkool en (rogge)brood. Daarnaast worden melkzuurbacteriën op grote schaal gebruikt in zogenaamde probiotica: drankjes en yoghurtjes die onze darmflora verbeteren. Voorbeelden van drankjes met “gezonde” melkzuurbacteriën zijn Yakult en Vifit. Het mag duidelijk zijn dat deze kleine beestjes van groot belang zijn voor de levensmiddelenindustrie en dat onderzoek aan melkzuurbacteriën met grote interesse door deze industrie gevolgd wordt.

Het aanwezig zijn van melkzuurbacteriën in zoveel verschillende omgevingen (melk, planten en onze darm) toont aan dat ze enorm goed in staat zijn om te functioneren in en zich aan te passen aan verschillende omgevingen. Mijn promotieonderzoek spitste zich vooral toe op het beantwoorden van de vraag hoe melkzuurbacteriën dat doen. Als voorbeeld van een melkzuurbacterie die zich goed kan aanpassen hebben we gekozen voor *Lactobacillus plantarum*, een bacterie die gevonden wordt op planten, worst, maar ook gebruikt wordt als “gezonde” bacterie in een vruchtendrank in Scandinavië en in staat is in ons maag-darm kanaal te leven.

Dit onderzoek is gedaan op niveau van de genetica. Genetica is de studie naar de functie van genen. Genen zijn de blauwdrukken voor alle functionele eigenschappen die levende wezens (planten, dieren, maar ook bacteriën) bezitten. Deze blauwdrukken noemen we de genetische code. Genen die je bezit maken wie je bent en hoe je er uit ziet. Genen zitten samen met elkaar op een groot molecuul, het DNA, waarvan er een of meerdere aanwezig zijn in iedere cel. Die complete verzameling DNA moleculen, met daarop de genen, noemt men het genoom. Het genoom van een bacterie bevat alle informatie die het nodig heeft om in alle mogelijke situaties te overleven. Genen worden door een proces wat men “transcriptie en translatie” noemt omgezet in eiwitten. Deze eiwitten zijn de uiteindelijke uitvoerders van de boodschap van het gen. Eiwitten zijn er in vele verschillende soorten waaronder voedselverwerkende eiwitten (enzymen), die zorgen voor energie en reparatie-eiwitten, die schade herstellen aan verschillende moleculen in de cel. Echter, een bacterie heeft niet te allen tijde al die functies nodig. Er zijn dus systemen aanwezig (ook gecodeerd door genen) om die functies alleen “aan” te zetten op het moment dat ze nodig zijn. Die controle-eiwitten noemt men regulators en zijn verantwoordelijk voor het aan dan wel uit zetten van de “transcriptie en translatie”. Bestudering van die eiwitten leidt tot inzicht in hoe een bacterie zich aanpast op het moment dat zijn omgeving verandert.

In hoofdstuk 2 van dit proefschrift wordt beschreven dat *L. plantarum* een veel groter genoom heeft (veel meer genen, eigenschappen) dan veel andere melkzuurbacteriën.

Dit heeft te maken met de verscheidenheid aan milieus waarin *L. plantarum* wordt gevonden (worst, zuurkool, darm) in vergelijking met andere melkzuurbacterien (die bijvoorbeeld alleen maar in melk voorkomen). Verder zien we dat het aantal regulators dat betrokken is bij het “aanschakelen” van functies onevenredig toeneemt met de grootte van het genoom. Kort gezegd, als je meer functies in je genoom hebt gecodeerd moet je naar verhouding steeds meer moeite stoppen in het aan- en uitschakelen van die functies. De verdere hoofstukken van het proefschrift richten zich op het vinden van de schakelaars op het genoom. Deze schakelaars liggen op het DNA in de buurt van een gen en worden herkend door de regulators die vervolgens de schakelaar kunnen activeren. In Hoofdstuk 3, 6 en 7 zijn er verschillende computermethodes gebruikt om deze schakelaars te vinden. Als basis werd gebruik gemaakt van algorithmes die zoeken naar patronen in het DNA die meer dan gemiddeld in deze, op zich willekeurige, gebieden voorkomen. Van genen die in de buurt van vergelijkbare patronen op het DNA aanwezig zijn wordt dan aangenomen dat die op dezelfde momenten aan staan. Deze bevindingen worden getoetst door gebruik te maken van data uit het laboratorium waarbij onder verschillende omstandigheden beschreven wordt welke genen er op dat moment wel en welke niet aan staan. In de hoofdstukken 4 en 5 worden een aantal schakelaars beschreven die werken zonder de tussenkomst van een regulator, zogenaamde RNA switches. Deze switches worden direct geactiveerd door een signaal in de cel (bijvoorbeeld een overschot aan bouwstoffen) doordat de bouwstof zelf de schakelaar aan zet. Deze nog niet zo bekende groep van schakelaars blijkt enorm subtiel en slim uitgedacht en zijn misschien meer te vergelijken met een dimmer dan een normale aan/uit schakelaar.

Alle informatie over de aanpassing van *L. plantarum* aan zijn nieuwe omgeving leidt uiteindelijk tot een complex netwerk (het regulatoire netwerk) van genen die samen of door elkaar beinvloedt worden. Dit netwerk helpt onderzoekers verder te begrijpen waarom er iets verandert in de activiteit van een bepaalde functie wanneer men die van een andere verandert.

Dankwoord

Het feit dat ik op verschillende werkplekken heb gezeten gedurende mijn promotie betekent ook dat er veel mensen betrokken zijn geweest bij het tot stand komen van mijn proefschrift. Omdat ik een, nogal, vergeetachtig karakter heb, weet ik bijna zeker dat ik daarom mensen ga vergeten. Bij deze wil ik dus iedereen die niet bij naam wordt genoemd, maar toch een bijdrage aan dit proefschrift heeft geleverd, bedanken voor hun inzet.

Als eerste wil ik mijn (co-)promotoren Roland, Michiel en Willem bedanken. Niet iedereen heeft de luxe om maar liefst drie hoogleraren te hebben die richting aan je onderzoek geven. Ik wil jullie dan ook bedanken voor de inbreng die jullie hebben geleverd. De verschillen in jullie karakters maakte dat ik op elk moment de juiste stimulans kreeg om mijn onderzoek door te zetten. Roland, als begeleider waar je op zowat elk moment binnen kunt vallen om te praten over alle zaken die te maken hebben met werk of daarbuiten. Michiel als grote motivator, of het werk nu mee of tegenzit, na een overleg met jou ben ik altijd weer vol goede moed aan het werk gegaan. Willem, alhoewel jij vooral op de achtergrond actief bent geweest verbaasde ik mij er altijd over hoe je (tot in groot detail) toch altijd wist waar ik mee bezig was.

Achter deze “grote drie” moet ik eigenlijk in een adem ook Christof noemen. Alhoewel jij pas na de eerste twee jaar van mijn promotie bij het CMBI binnen kwam, heb ik in die twee jaar veel aan jouw input gehad. Het feit dat je mede-auteur bent in drie van de zes hoofstukken geeft dit goed weer. Naast mijn begeleiders wil ik ook alle BAMICS-ers bedanken voor alle gezelligheid. Met name mijn vaste “roomies”: Anita, Mengjin, Miao Miao, Mark, Robert en Jos en alle studenten die een tijdelijke intrek op “onze” kamer hebben gehad. Bedankt voor de leuke uren, zowel tijdens werktijd als na half zes (Liero, Midtown Madness). In het bijzonder wil ik mijn studenten Wendy, Anke, Tom, Lex, Joris en Daniel bedanken die als stagiaire bij mij allemaal een wezenlijke bijdrage aan dit proefschrift hebben geleverd. Ook onze (verre) buren Richard P., Juma en Barzhan: dank voor alle interessante wetenschappelijke als niet-wetenschappelijke discussies. Ook de postdocs die in de loop der tijd zijn verschenen (en sommige ook weer verdwenen), Bernadet, Bas, Frank, Peter en Greer (Wilson, S.C.) dank voor jullie input. Jos, Victor en Wilmar bedankt voor het draaiend houden van de verschillende bamics-servers en het installeren van de vele verschillende tooltjes en databases die ik nodig had voor mijn werk. Verder bedankt voor de gezellige BAMICS-etentjes, speurtochten en Wii avonden. Alle andere (ex-) CMBI-ers bedankt voor de plezierige tijd o.a. aan de koffietafel.

Naast mijn collega’s uit Nijmegen wil ik ook mijn NIZO collega’s bedanken. In het bijzonder mijn labmaatje van het eerste uur Roger. Dank voor je geduld en waardering voor mijn twee linkerhanden. Mede AIO-“lotgenoten” van de denktank, Jolanda, Mariela, Marc en Arno, ik kijk uit naar jullie boekjes, het wordt een gezellig eerste half jaar met al die feestjes! Douwe bedankt voor je wetenschappelijke input en de CDtjes met BASE dumps voor Correlex.

Mede WCFS-ers (of TIFNers) actief binnen C-007 die nog niet genoemd zijn (Peter, Bart, Maaike, Maria, Sonja, Freddy, Andrea, Carien, Erwin en alle andere begeleiders en experts): bedankt voor alle inbreng tijdens de projectmeetings. Cindy, bedankt voor de interesse en het omzeilen van alle mogelijke WCFS-beren op de weg. Ook de leden van de andere projecten met wie ik in de afgelopen vijf jaar heb samengewerkt, DANK! In het bijzonder wil ik nog Johan en Roy bedanken voor het bieden van een overbruggingsperiode naar mijn nieuwe baan op NIZO, omdat het mij in staat stelde om dicht bij het vuur te blijven, wat een goede motivatie is om de laatste eindjes aan elkaar te knopen. Niet in de laatste plaats omdat het werk dat ik in die tijd heb gedaan erg leuke resultaten heeft opgeleverd en waarschijnlijk tot een aantal mooie co-publicaties zal leiden. Naast de WCFS-ers wil ik ook alle mede projectleden van het IOP-Genomics en Kluyver Centre bedanken.

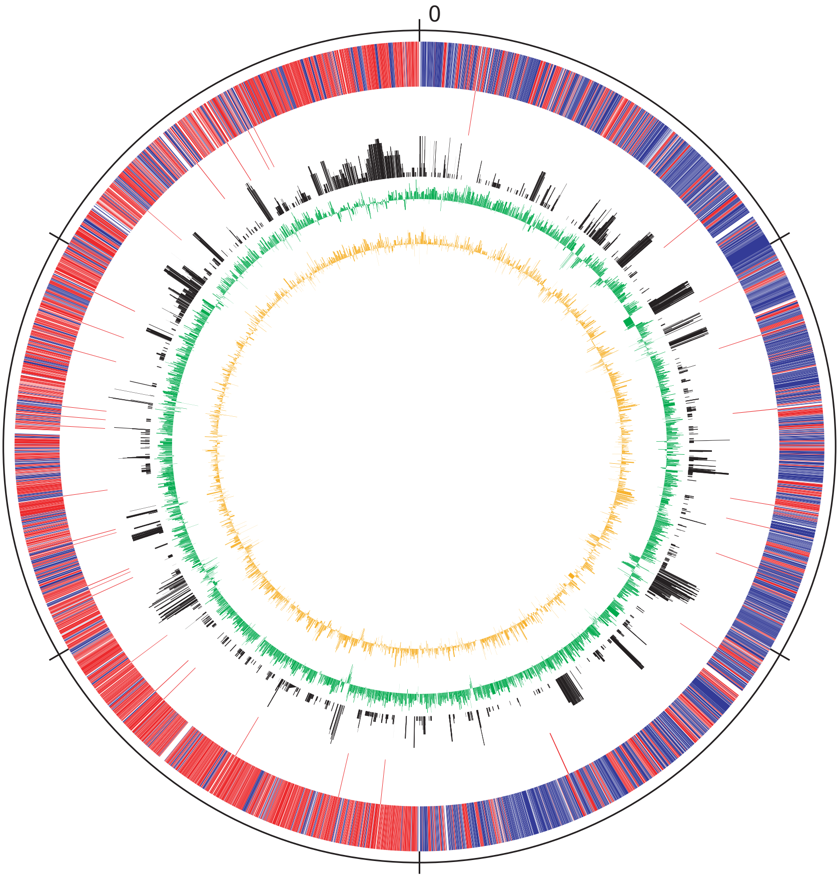
Als laatste wil ik ook mijn dank uitspreken naar familie en vrienden die voor de nodige afleiding zorgen na een week hard werken. In het bijzonder wil ik hierbij de trainingen en wedstrijden met mijn teamgenoten van Black Eagles Heren 6 noemen. Elk jaar worden we een stapje beter mannen, misschien dat we voor 2020 nog eens kampioen worden! Frank, bedankt voor de slappe klets vooral zo rond midden februari, langzaamaan beginnen ze hier boven de rivieren ook te realiseren dat carnaval meer is dan drie dagen bier zuipen. Ook mijn (schoon)familie, Sander, Frouke, Maas, Joke, Anouk en Jean-Paul: bedankt voor jullie interesse en luisterend oor, het is niet altijd makkelijk om naar zo’n saaie onderzoeker te luisteren. Pa en Ma, bedankt voor de vrijheid en steun die ik tijdens mijn hele opleiding van jullie heb gehad.

Als aller-aller-laatste wil ik mijn lieffie-lief bedanken: Sjarai, naast de tijd en moeite die je in het opmaken van het proefschrift hebt gestoken, bovenal bedankt voor je geduld, die vele vele zaterdagen, ik weet dat wachten niet in je karakter zit.

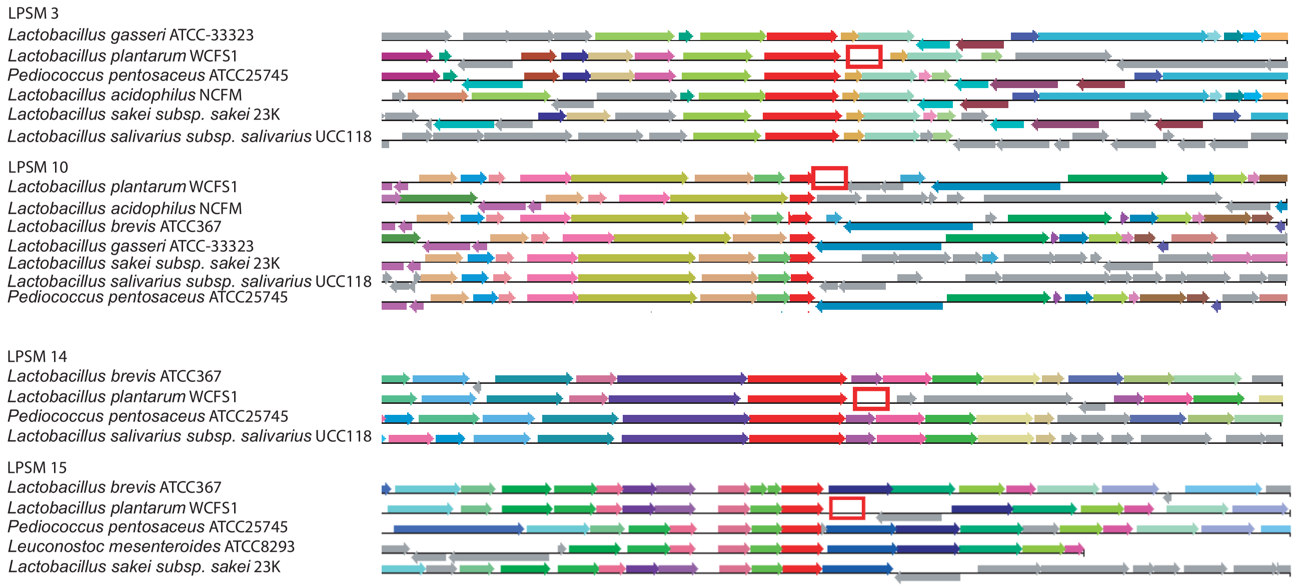
Michiel

Colour Figures

Chapter 4



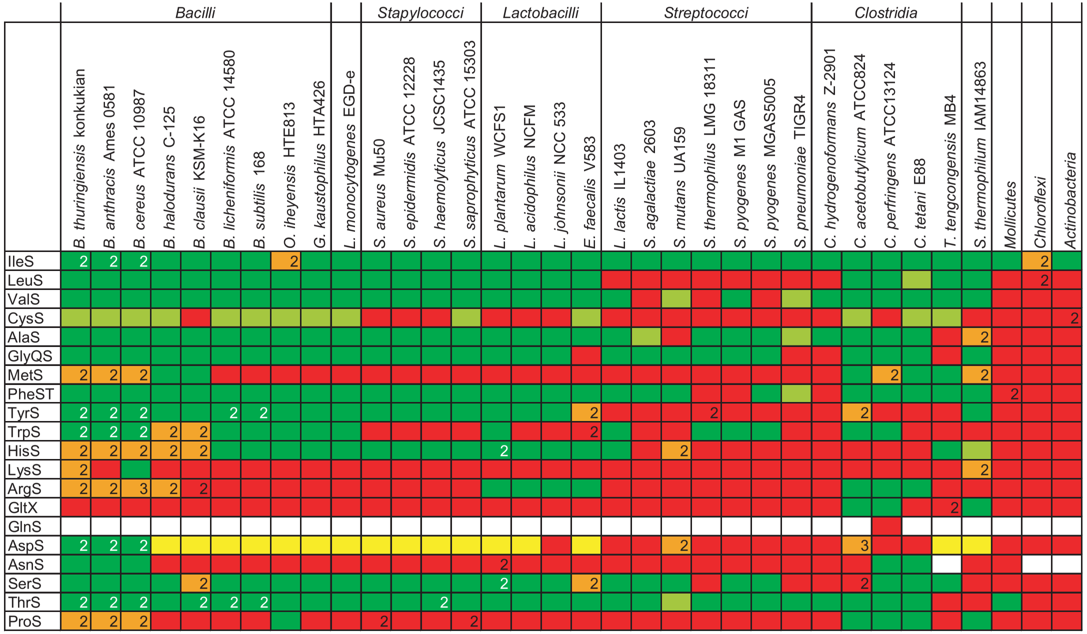
Chapter 4, Figure 1: Chromosome wheel displaying the identified LPSM hits on the chromosome of *L. plantarum*. Outer ring (ring 1): predicted ORFs (+ and – strand, blue and red, respectively), ring 2: LPSM occurrences, ring 3: Relative conservation among 20 L. plantarum strains as determined by array-based chromosome profiling. Peak height represents the number of strains, in which specific chromosomal regions of strain WCFS1 were scored as absent in other strains of L. plantarum (17), ring 4: G+C content centered around the median G+C content, ring 5: Codon Adaptation index. Figure generated using MGV (46).


LPSM 3
Lactobacillus gasseri ATCC-33323
Lactobacillus plantarum WCFS1
Pediococcus pentosaceus ATCC25745
Lactobacillus acidophilus NCFM
Lactobacillus sakei subsp. sakei 23K
Lactobacillus salivarius subsp. salivarius UCC118
LPSM 10
Lactobacillus plantarum WCFS1
Lactobacillus acidophilus NCFM
Lactobacillus brevis ATCC367
Lactobacillus gasseri ATCC-33323
Lactobacillus sakei subsp. sakei 23K
Lactobacillus salivarius subsp. salivarius UCC118
Pediococcus pentosaceus ATCC25745
LPSM 14
Lactobacillus brevis ATCC367
Lactobacillus plantarum WCFS1
Pediococcus pentosaceus ATCC25745
Lactobacillus salivarius subsp. salivarius UCC118
LPSM 15
Lactobacillus brevis ATCC367
Lactobacillus plantarum WCFS1
Pediococcus pentosaceus ATCC25745
Leuconostoc mesenteroides ATCC8293
Lactobacillus sakei subsp. sakei 23K

Chapter 4, Figure 2: Genome context of six LPSM locations in *L. plantarum* WCFS1 compared to other lactic acid bacteria.
LPSM position in *L. plantarum* is denoted by the red box. Orthologous genes in different organism have the same color. Figure generated using the ERGO Bioinformatics suite (47).

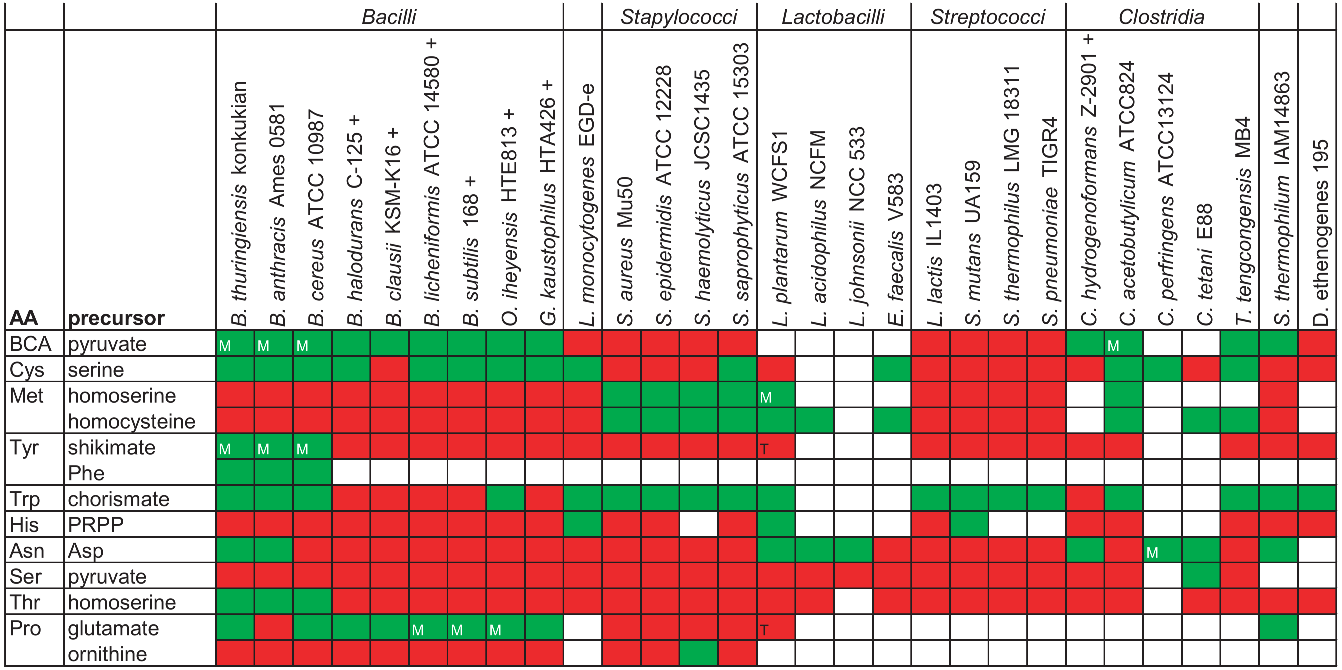
Colour Figures

Chapter 5, Part 1



	Bacilli										Stapylococci				Lactobacilli				Streptococci							Clostridia								
	B. thuringiensis konkukian	*B. anthracis* Ames 0581	*B. cereus* ATCC 10987	*B. halodurans* C-125	*B. clausii* KSM-K16	*B. licheniformis* ATCC 14580	*B. subtilis* 168	*O. iheyensis* HTE813	*G. kaustophilus* HTA426	*L. monocytogenes* EGD-e	*S. aureus* Mu50	*S. epidermidis* ATCC 12228	*S. haemolyticus* JCSC1435	*S. saprophyticus* ATCC 15303	*L. plantarum* WCFS1	*L. acidophilus* NCFM	*L. johnsonii* NCC 533	*E. faecalis* V583	*L. lactis* IL1403	*S. agalactiae* 2603	*S. mutans* UA159	*S. thermophilus* LMG 18311	*S. pyogenes* M1 GAS	*S. pyogenes* MGAS5005	*S. pneumoniae* TIGR4	*C. hydrogenoformans* Z-2901	*C. acetobutylicum* ATCC824	*C. perfringens* ATCC13124	*C. tetani* E88	*T. tengcongensis* MB4	*S. thermophilum* IAM14863	*Mollicutes*	*Chloroflexi*	*Actinobacteria*
IleS	🟩 2	🟩 2	🟩 2	🟩	🟩	🟩	🟩	🟧 2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟧 2	🟩
LeuS	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	▨	🟩	🟩	🟥	🟥 2	🟥
ValS	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	▨	🟥	🟩	🟥	▨	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥
CysS	▨	▨	▨	▨	🟥	▨	▨	▨	▨	▨	🟥	🟥	🟥	▨	🟥	🟥	🟥	▨	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	▨	🟥	▨	▨	🟥	🟥	🟥	🟥 2
AlaS	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	▨	🟥	🟩	🟩	🟩	▨	🟩	🟩	🟩	🟩	🟥	🟧 2	🟥	🟥	🟥
GlyQS	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥
MetS	🟧 2	🟧 2	🟧 2	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟧 2	🟩	🟩	🟧 2	🟥	🟥	🟥
PheST	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟩	▨	🟥	🟩	🟩	🟩	🟩	🟩	🟥 2	🟥	🟥
TyrS	🟩 2	🟩 2	🟩 2	🟩	🟩	🟩 2	🟩 2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟧 2	🟥	🟥	🟥	🟥 2	🟥	🟥	🟥	🟥	🟧 2	🟥	🟥	🟥	🟩	🟥	🟥	🟥
TrpS	🟩 2	🟩 2	🟩 2	🟧 2	🟧 2	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥 2	🟩	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥
HisS	🟧 2	🟧 2	🟧 2	🟧 2	🟧 2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩 2	🟩	🟩	🟩	🟩	🟥	🟧 2	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	▨	🟥	🟥	🟥
LysS	🟧 2	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟧 2	🟥	🟥	🟥
ArgS	🟧 2	🟧 2	🟧 3	🟧 2	🟥 2	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥
GltX	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟥 2	🟩	🟥	🟥	🟥
GlnS	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟥	⬜	⬜	⬜	⬜	⬜	⬜
AspS	🟩 2	🟩 2	🟩 2	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟥	🟨	🟥	🟥	🟧 2	🟥	🟥	🟥	🟥	🟥	🟧 3	🟥	🟥	🟨	🟨	🟥	🟥	🟥
AsnS	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥 2	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	⬜	🟥	🟥	⬜	⬜
SerS	🟩	🟩	🟩	🟩	🟧 2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩 2	🟩	🟩	🟧 2	🟩	🟩	🟩	🟥	🟩	🟩	🟥	🟥	🟥 2	🟩	🟩	🟩	🟥	🟥	🟥	🟥
ThrS	🟩 2	🟩 2	🟩 2	🟩	🟩 2	🟩 2	🟩 2	🟩	🟩	🟩	🟩	🟩	🟩 2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	▨	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟥	🟥
ProS	🟧 2	🟧 2	🟧 2	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥 2	🟥	🟥	🟥 2	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥

Chapter 5, Figure 2: T-box regulation of tRNA ligase encoding genes in the *Firmicutes*.
The color coding relates to the presence or absence of a T-box upstream of the genes encoding the amino acid-specific tRNA ligases in the various species and strains. Green indicates the tRNA ligase(s) is (are) regulated by a T-box and red that the tRNA ligase(s) is (are) not regulated by a T-box. Although most tRNA ligases are present in one copy on the genome, several organisms contain two, or in some cases three copies of specific ligases (indicated by a number in the box). Orange indicates that 1 of the 2 tRNA ligases is regulated by a T-box or 1 out of 3 in the case of the *argS* genes in *B. cereus* ATCC 10987 and the aspS genes in *C. acetobutylicum*. Light green indicates that the tRNA ligase is not the first in the operon, but is regulated by a T-box with the same specificity. Yellow color coding indicates that the regulated tRNA ligase is the second gene in an operon in combination with another tRNA ligase gene. These genes are probably regulated by a T-box with different specificity than the specificity of the tRNA ligase. White indicates that no tRNA ligase of this type present in the organism. In principle, a species needs at least one specific tRNA ligase for each amino acid. Nevertheless, there are exceptions. For instance, all but one *(Clostridium perfringens)* of the analyzed genomes, lack the gene that encodes a Gln-tRNA ligase and the genomes of the *Chloroflexi, Actinobacteria* and *Thermoanaerobacter tencongens* also lack an Asn-tRNA ligase. In these cases, the biological role of the Gln-tRNA ligase is taken over by the Glu tRNA ligase, which couples a Glu residue to the tRNAGln. The residue is subsequently transformed into a Gln by a tRNA specific amidotransferase (50). Similarly, an Asn-tRNAAsn is formed via transamidation of an Asp residue (Asp-tRNAAsn to Asn-tRNAAsn) in bacteria that lack an Asn tRNA ligase (51). Consequently, we found that all species lacking either the Gln-tRNA ligase or the Asn-tRNA ligase have an orthologous gene coding for the corresponding amidotransferase. No T-boxes were identified upstream of those genes.

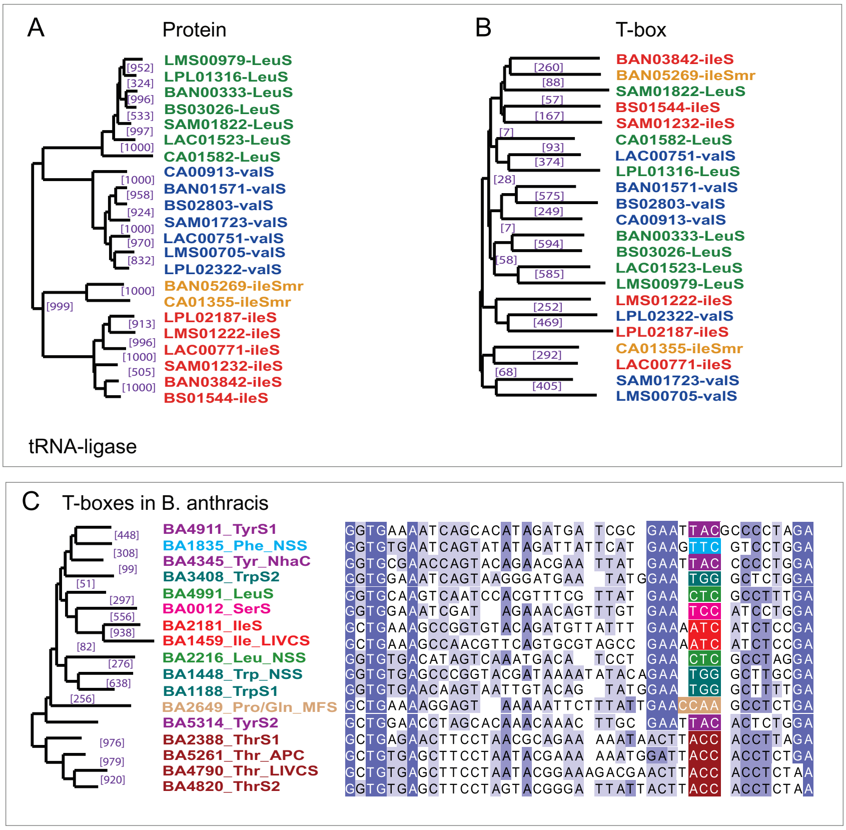


		Bacilli										*Stapylococci*				*Lactobacilli*				*Streptococci*				*Clostridia*						
AA	**precursor**	*B. thuringiensis* konkukian	*B. anthracis* Ames 0581	*B. cereus* ATCC 10987	*B. halodurans* C-125 +	*B. clausii* KSM-K16 +	*B. licheniformis* ATCC 14580 +	*B. subtilis* 168 +	*O. iheyensis* HTE813 +	*G. kaustophilus* HTA426 +	*L. monocytogenes* EGD-e	*S. aureus* Mu50	*S. epidermidis* ATCC 12228	*S. haemolyticus* JCSC1435	*S. saprophyticus* ATCC 15303	*L. plantarum* WCFS1	*L. acidophilus* NCFM	*L. johnsonii* NCC 533	*E. faecalis* V583	*L. lactis* IL1403	*S. mutans* UA159	*S. thermophilus* LMG 18311	*S. pneumoniae* TIGR4	*C. hydrogenoformans* Z-2901 +	*C. acetobutylicum* ATCC824	*C. perfringens* ATCC13124	*C. tetani* E88	*T. tengcongensis* MB4	*S. thermophilum* IAM14863	*D. ethenogenes* 195
BCA	pyruvate	🟩 M	🟩 M	🟩 M	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥					🟥	🟥	🟥	🟥	🟩	🟩 M			🟩	🟩	🟥
Cys	serine	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟥			🟩	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟥	🟩	🟥	🟥
Met	homoserine	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩 M				🟥	🟥	🟥	🟥		🟩				🟥	
	homocysteine	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟥	🟥	🟥	🟥		🟩		🟩	🟩	🟥	
Tyr	shikimate	🟩 M	🟩 M	🟩 M	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥 T				🟥	🟥	🟥	🟥	🟥	🟥			🟥	🟥	🟥
	Phe	🟩	🟩	🟩																										
Trp	chorismate	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩				🟩	🟩	🟩	🟩	🟥	🟩			🟩	🟩	🟩
His	PRPP	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥		🟥	🟩				🟥	🟩			🟥	🟥			🟥	🟥	🟥
Asn	Asp	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩 M	🟩	🟥	🟩	
Ser	pyruvate	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥		🟩	🟥		
Thr	homoserine	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥	🟥	🟥
Pro	glutamate	🟩	🟥	🟩	🟩	🟩	🟩 M	🟩 M	🟩 M	🟩		🟥	🟥	🟥	🟥	🟥 T													🟩	
	ornithine	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥		🟥	🟥	🟩	🟥															

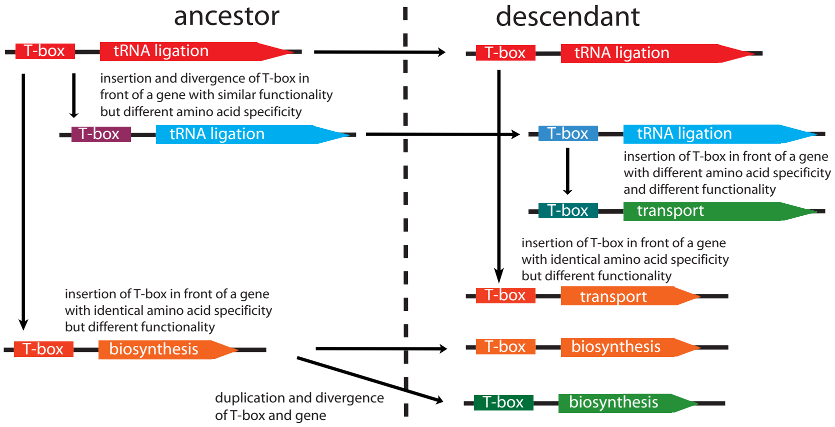
Chapter 5, Figure 3: T-boxes preceding the genes related to amino acid biosynthesis in *Firmicutes*.
Color coding identifies the presence of the biosynthesis pathway and whether it is regulated by a T-box: Green; T-box regulated; red; not T-box regulated; no color; pathway absent. +TRAP protein is present. M Pathway genes organized in multiple operons. BCA indicates the branched chain amino acids valine, leucine and isoleucine.

Colour Figures

Chapter 5, Part 2


A
Protein
LMS00979-LeuS
LPL01316-LeuS
BAN00333-LeuS
BS03026-LeuS
SAM01822-LeuS
LAC01523-LeuS
CA01582-LeuS
CA00913-valS
BAN01571-valS
BS02803-valS
SAM01723-valS
LAC00751-valS
LMS00705-valS
LPL02322-valS
BAN05269-ileSmr
CA01355-ileSmr
LPL02187-ileS
LMS01222-ileS
LAC00771-ileS
SAM01232-ileS
BAN03842-ileS
BS01544-ileS
[952]
[324]
[996]
[533]
[997]
[1000]
[1000]
[958]
[924]
[1000]
[970]
[832]
[1000]
[999]
[913]
[996]
[1000]
[505]
[1000]
tRNA-ligase
B
T-box
BAN03842-ileS
BAN05269-ileSmr
SAM01822-LeuS
BS01544-ileS
SAM01232-ileS
CA01582-LeuS
LAC00751-valS
LPL01316-LeuS
BAN01571-valS
BS02803-valS
CA00913-valS
BAN00333-LeuS
BS03026-LeuS
LAC01523-LeuS
LMS00979-LeuS
LMS01222-ileS
LPL02322-valS
LPL02187-ileS
CA01355-ileSmr
LAC00771-ileS
SAM01723-valS
LMS00705-valS
[260]
[88]
[57]
[167]
[7]
[93]
[374]
[28]
[575]
[249]
[7]
[594]
[58]
[585]
[252]
[469]
[292]
[68]
[405]
C
T-boxes in B. anthracis
BA4911_TyrS1 GGTGAAAATCAGCACATAGATGA TCGC GAATTACGCCCTAGA
BA1835_Phe_NSS GGTGTGAATCAGTATATAGATTATTCAT GAAGTTC GTCCTGGA
BA4345_Tyr_NhaC GGTGCGAACCAGTACAGAACGAA TTAT GAATTAC CCCCTGGA
BA3408_TrpS2 GGTGGAAATCAGTAAGGGATGAA TATGGAA TGG GCTCTGGA
BA4991_LeuS GGTGCAAGTCAATCCACGTTTCG TTAT GAA CTC GCCTTTGA
BA0012_SerS GGTGGAAATCGAT AGAAACAGTTTGT GAA TCC ATCCTGGA
BA2181_IleS GCTGAAAGCCGGTGTACAGATGTTATTT GAAAATC ATCTCCGA
BA1459_Ile_LIVCS GCTGAAAGCCAACGTTCAGTGCGTAGCC GAAAATC ATCTCCGA
BA2216_Leu_NSS GGTGTGACATAGTCAAATGACA TCCT GAA CTC GCCTAGGA
BA1448_Trp_NSS GGTGTGAGCCCGGTACGATAAAATATACAGAA TGG GCTTGCGA
BA1188_TrpS1 GGTGTGAACAAGTAATTGTACAG TATGGAA TGG GCTTTTGA
BA2649_Pro/Gln_MFS GCTGAAAAGGAGT AAAAATTCTTTATTGAACCAA GCCTCTGA
BA5314_TyrS2 GCTGGAACCTAGCACAAACAAAC TTGC GAATTAC ACTCTGGA
BA2388_ThrS1 GCTGAGAACTTCCTAACGCAGAA AAATAACTTACC ACCTTAGA
BA5261_Thr_APC GCTGTGAGCTTCCTAATACGAAA AAATGGATTACC ACCTCTGA
BA4790_Thr_LIVCS GCTGTGAGCTTCCTAATACGGAAAGACGAACTTACC ACCTCTAA
BA4820_ThrS2 GGTGTGAGCTTCCTAGTACGGGA TTATTACTTACC ACCTCTAA
[448]
[308]
[99]
[51]
[297]
[556]
[938]
[82]
[276]
[638]
[256]
[976]
[979]
[920]

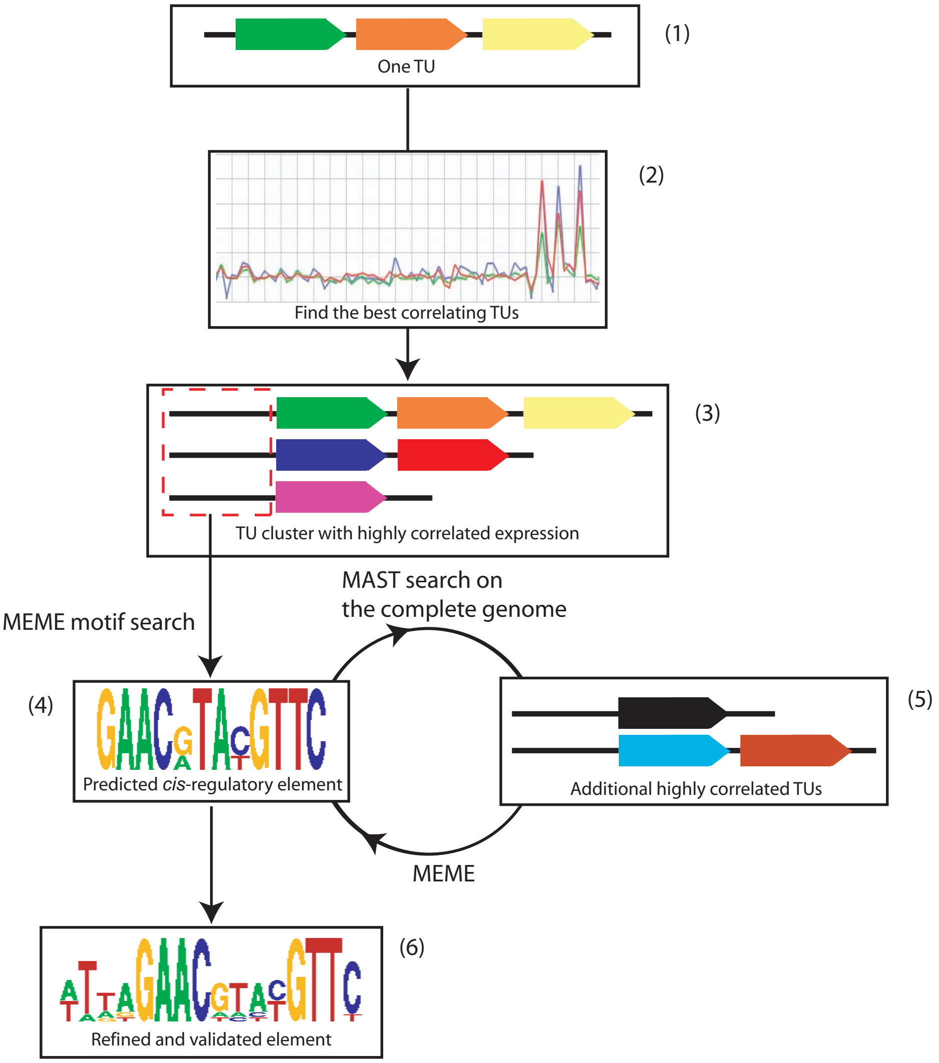
Chapter 5, Figure 4: The evolutionary relationship between some T-boxes.
(A) shows a putative phylogeny of the branched-chain amino acid tRNA ligases of *B. anthracis* Ames, *B. subtilis* 168, *C. acetobutylicum* ATCC824D, *L. acidophilus* NCFM, *L. plantarum* WCFS1, *L. mesenteroides* ATCC8293 and *S. aureus* Mu50. (B) shows the Neighbor Joining tree for the related T-boxes. This tree does not reflect the true phylogeny of the regulatory elements but merely serves as an indicator of element similarity. (C) shows the Neighbor Joining tree for various T-boxes found in *B. anthracis* Ames next to the specifier codon containing part of the corresponding multiple sequence alignment. The specifier codon is indicated in white letters. The amino acid specificity is color-coded: Red and orange relate to Ile, green to Leu, light blue to Phe, beige to Pro or Gln, pink to Ser, brown to Thr, turquoise to Trp, purple to Tyr and dark blue to Val. The functional group of the regulated gene is indicated by the letters that follow the amino acid code. The group can be either a transporter of the APC, LIVCS, MFS, NhaC or NSS family or a tRNA-ligase (S or Smr for mupirocin-resistant tRNA ligase).The NSS-family transport proteins regulated by a Leu, Phe and Trp T-box are in-paralogs characteristic for the species of the *Bacillus cereus* group. The purple numbers between brackets indicate the bootstrap support for the displayed clusters in the Neighbor Joining trees (out of 1000).


ancestor
descendant
T-box
tRNA ligation
insertion and divergence of T-box in front of a gene with similar functionality but different amino acid specificity
T-box
tRNA ligation
insertion of T-box in front of a gene with identical amino acid specificity but different functionality
T-box
biosynthesis
duplication and divergence of T-box and gene
T-box
tRNA ligation
T-box
tRNA ligation
insertion of T-box in front of a gene with different amino acid specificity and different functionality
T-box
transport
insertion of T-box in front of a gene with identical amino acid specificity but different functionality
T-box
transport
T-box
biosynthesis
T-box
biosynthesis

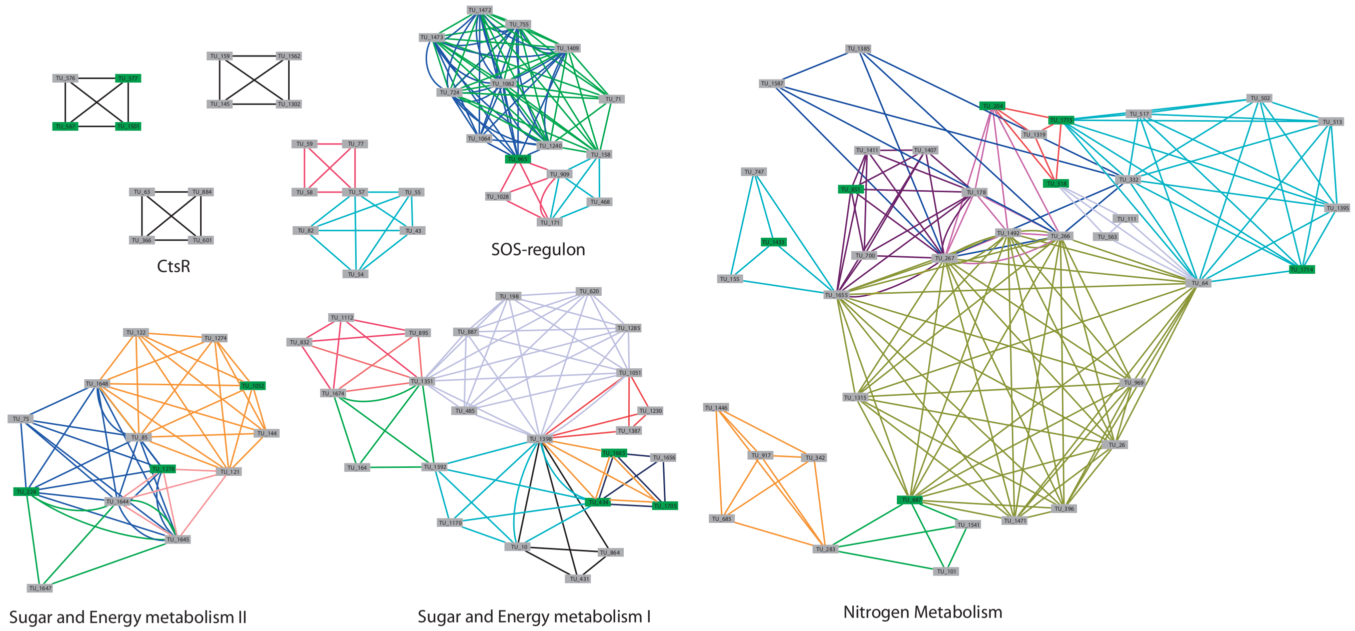
Chapter 5, Figure 5: Scenarios for T-box evolution.
Most T-boxes are linked on the genome to genes that encode proteins related to three main functional categories, namely: tRNA-ligation, amino acid transport and amino acid biosynthesis. Our results suggest that there are several ways in which these T-boxes (amino acid specificity is color-coded) evolve: T-boxes are duplicated and inserted in front of genes from the same functional category but with different amino acid specificity and, vice versa, in front of genes with the same amino acid specificity but from another functional category. In addition, duplication and insertion in front of genes from another functional category with different amino acid specificity takes place. These different routes impose convergent as well as divergent constraints on the evolution of the T-box sequence. As shown, already within a few speciations a plethora of T-box sequences could develop from one initial sequence. In effect, these processes assure a more or less independent evolution of the T-boxes from the genes they control.

Colour Figures

Chapter 7


(1)
One TU
(2)
Find the best correlating TUs
(3)
TU cluster with highly correlated expression
MAST search on
the complete genome
MEME motif search
(4)
Predicted *cis*-regulatory element
(5)
Additional highly correlated TUs
MEME
(6)
Refined and validated element

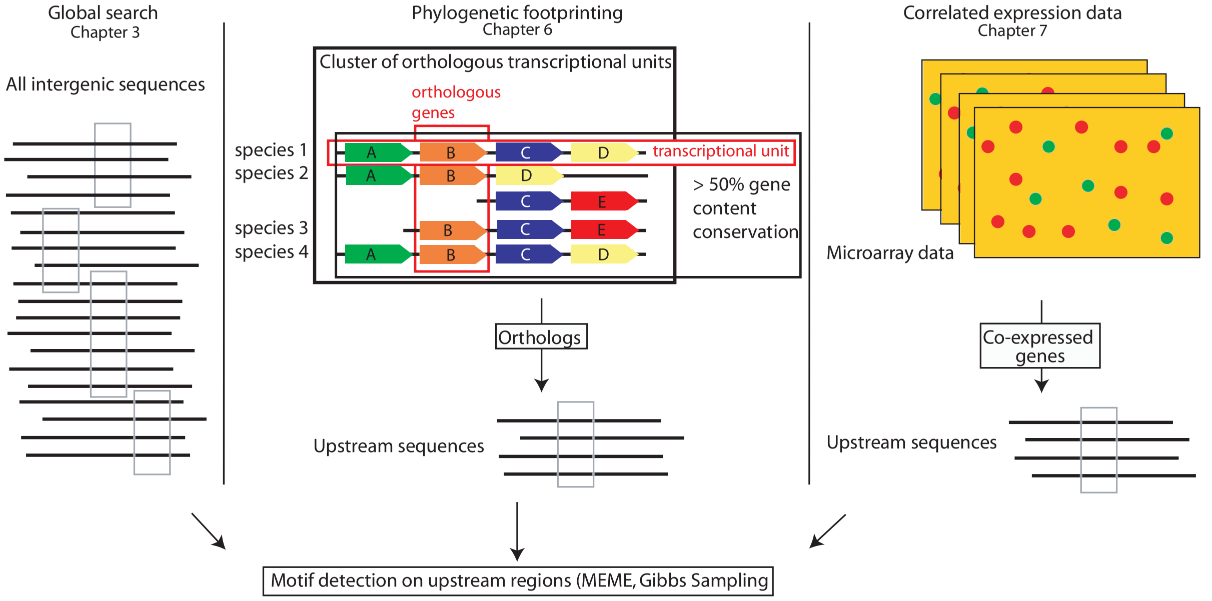
Chaper 7, Figure 1: Flowchart of the followed procedure.
1= query TU, 2= identification of co-regulated TUs in a large set of experiments, 3= selection of upstream sequences of co-regulated TU cluster, 4= shared cis-element detection, 5= iteration procedure to expand or complete regulon, 6= refined regulon cis-element



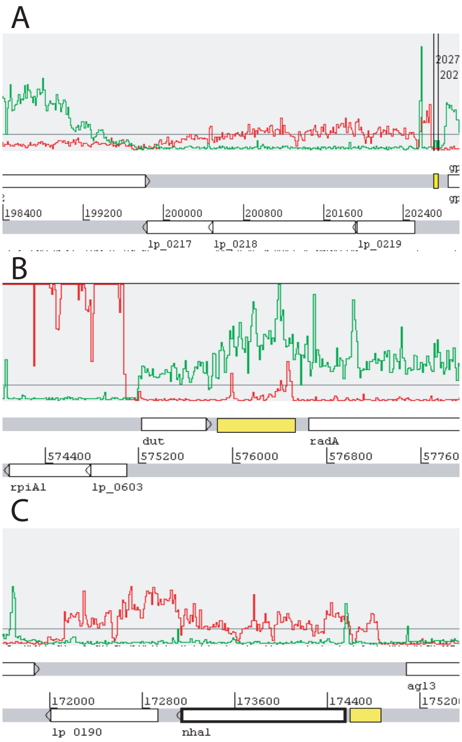
Chapter 7, Figure 3: Regulatory network of *L. plantarum* based on iteratively expanded regulons.
Nodes represent different TUs; lines connect TUs that are member of the same regulon, distinguished by different colors per regulon. TUs colored in green contain at least one gene encoding a regulatory protein. A detailed view on the network can be found at www.cmbi.ru.nl/~mwels/Chapter_7.

Colour Figures

Chapter 8



Chapter 8, Figure 1: different procedures used to predict regulatory elements in *L. plantarum*



Chapter 8, Figure 3: Different examples of the use of tiling microarray data to validate transcriptome predictions. Expression levels are shown for the + (green) and – (red) strand. A) shows the start of transcription of a TU approximately 200 nt upstream of the first gene of the TU (*lp_0219*), near the predicted promoter (202705 – 202746, indicated in yellow) and continues to the end of the TU (*lp_0217*). B) shows the transcription of one of the LPSMs (indicated in yellow) as being part of a transcript that starts upstream of dut and continues expression into radA. C) shows the transcription of a T-box (indicated in yellow) in the upstream region of *lp_0191* (annotated as a natrium antiporter (Nha1) in the original *L. plantarum* annotation).

Curriculum Vitae

Michiel Wouter Wilhelmus Wels werd geboren op 16 februari 1980 te ’s-Hertogenbosch. In 1984 verhuisde het gezin naar het naburige Rosmalen, waar hij basisschool ’t Ven en het VWO (Rodenborch College) bezocht. In 1998 behaalde hij zijn diploma voor het Atheneum en vervolgde zijn opleiding aan de Radboud Universiteit Nijmegen (voorheen Katholieke Universiteit Nijmegen). Aan die universiteit begon hij met een opleiding Biologie en behaalde in 2002 zijn diploma. Tijdens zijn opleiding deed Michiel onderzoek aan het transport van hormonen in de cellen van de hypofyse van de Afrikaanse klauwkikker (*Xenopus laevis*). Dit onderzoek werd verricht aan de vakgroep Moleculaire Dierfysiologie, onder begeleiding van Prof. Dr. Gerard J.M. Martens en Dr. Gerrit Bouw. Na deze periode werd zijn interesse gewekt voor het veld van de bioinformatica en begon Michiel een onderzoek naar geconserveerde gen clusters in *Lactobacillus plantarum* aan het Center for Molecular and Biomolecular Informatics (CMBI) bij de vakgroep “Bacterial Genomics” onder leiding van Prof. Dr. Roland J. Siezen en Dr. Jos Boekhorst. In 2002 werd hem vanuit die vakgroep de mogelijkheid geboden om een promotieonderzoek te doen naar de transcriptie regulatie van het genoom van *L. plantarum*. Alhoewel zijn onderzoek werd gedaan onder de vlag van de leerstoelgroep Microbiologie en onderdeel uitmaakte van het onderzoeksprogramma “Microbial Functionality and Food Safety” van het Top Institute Food and Nutrition (voorheen Wageningen Centre for Food Sciences), uitgevoerd onder de paraplu van de onderzoeksschool VLAG, waren zijn werklocaties het CMBI aan de Radboud Universiteit Nijmegen te Nijmegen en NIZO food research te Ede. Dit promotieonderzoek werd uitgevoerd onder de begeleiding van Prof. Dr. Willem M. de Vos, Prof. Dr. Roland J. Siezen, Prof. Dr. Michiel Kleerebezem en in een later stadium Dr. Christof Francke. Van 15 oktober 2006 tot 1 augustus 2007 was hij werkzaam als onderzoeker via Nutri-Akt BV voor het Top Instituut Food and Nutrition te Wageningen. Vanaf 1 augustus 2007 werkt hij als Scientist/Project leader voor NIZO food research te Ede en als PostDoc binnen het Top Institute Food and Nutrition.

List of publications

Wels M, Francke C, Kerkhoven R, Kleerebezem M, Siezen RJ
Predicting cis-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res. 2006 Apr 13;34(7):1947-58.*

Wels M, Groot Kormelink T, Kleerebezem M, Siezen RJ, Francke C
T-boxes and T-box regulated genes in bacteria: an in silico analysis. *Submitted for publication*

Wels M, Bongers R, Boekhorst J, Sturme M, Molenaar D, de Vos WM, Siezen RJ, Kleerebezem M
Large intergenic cruciform supermotifs in the *Lactobacillus plantarum* genome: a putative RNAi-like regulatory role? *Submitted for publication*

Wels M, Overmars L, Francke C, Kleerebezem M, Siezen RJ
Regulon prediction in *Lactobacillus plantarum* WCFS1 on basis of correlated microarray data. *Manuscript in preparation*

Wels M, Kleerebezem M, Siezen RJ
In silico promoter analysis in the intergenic regions of *Lactobacillus plantarum* WCFS1. *Manuscript in preparation*

Boekhorst J, Wels M, Kleerebezem M, Siezen RJ
The predicted secretome of *Lactobacillus plantarum* WCFS1 sheds light on interactions with its environment. Microbiology. *2006 Nov;152(Pt 11):3175-83.*

Serrano LM, Molenaar D, Wels M, Teusink B, Bron PA, de Vos WM, Smid EJ
Thioredoxin reductase is a key factor in the oxidative stress response of *Lactobacillus plantarum* WCFS1. *Microb Cell Fact. 2007 Aug 28;6(1):29.*

Francke C, Kerkhoven R, Wels M, Siezen RJ
Identification of LacI-specific operators to link local and global transcription factor activity. *Submitted for publication*

Tzeneva VA, Castioni A, Wels M, Phan HTK, Rademaker JLW, Siezen RJ, Starrenburg MJC, Torriani S, Smidt H, Molenaar D, van Hylckama Vlieg JET
Genomic and phenotypic diversity of *Lactobacillus plantarum* strains isolated from diverse environments. *Manuscript in preparation*

VLAG graduate school activities

Discipline specific activities
Spring school bioinformatics (2004, EPS)
Genetics and physiology of food-associated microorganisms (2004 ,VLAG)
Principles of ~omics data analysis (2006, EPS/VLAG)

European Conference on Bioinformatics, Saarbrucken, Germany (2002)
European Conference on Prokaryotic genomics, Gottingen Germany (2003, poster)
Bacterial Neural Networks, San Feliu de Guixols, Spain (2004, poster)

Conference on Gram-Positive Genomics, San Diego, USA (2005, oral presentation)
8th international conference on Lactic Acid Bacteria
(LAB8, volunteer in the organisation, 2006)
Dutch/Benelux bioinformatics conference (2003-4-5-6-7, oral presentation in 2007)

General courses
Scientific writing (UTN, Radboud University, 2004)

Presentation course (UTN, Radboud University, 2004)

Writing to be read (WCFS, 2003)
Debating course (WCFS, 2006)

Optional courses and activities
Preparation PhD research proposal

Regular group meetings (WCFS, NIZO, CMBI, IOP, Kluyver Centre) (2002 – 2007)
Attendance WCFS Food Summit (2003)
Attendance WCFS WE-days (excursions to Cork, Maastricht, Zeeland) (2003-2005)

Cover Design en Lay-out:
Sjarai van Butselaar en Michiel Wels