microbial biotechnology

Microbial Biotechnology (2010) **4**(3), 333–344 doi:10.1111/j.1751-7915.2010.00217.x

Reconstruction of the regulatory network of *Lactobacillus plantarum* WCFS1 on basis of correlated gene expression and conserved regulatory motifs

Michiel Wels,[1,2,3]* Lex Overmars,[1,3] Christof Francke,[1,3] Michiel Kleerebezem[1,2,4] and Roland J. Siezen[1,2,3]

[1]*Top Institute Food and Nutrition and Kluyver Centre for Genomics of Industrial Fermentation, PO Box 557, 6700 AN Wageningen, The Netherlands.*

[2]*NIZO Food Research, PO Box 20, 6710 BA Ede, The Netherlands.*

[3]*Radboud University Nijmegen Centre for Molecular and Biomolecular Informatics, PO Box 9101, 6500 HB Nijmegen, The Netherlands.*

[4]*Wageningen University and Research Centre, Dreijenplein 10, 6703 HB Wageningen, The Netherlands.*

Summary

Gene regulatory networks can be reconstructed by combining transcriptome data from many different experiments to elucidate relations between the activity of certain transcription factors and the genes they control. To obtain insight in the regulatory network of *Lactobacillus plantarum*, microarray transcriptome data from more than 70 different experimental conditions were combined and the expression profiles of the transcriptional units (TUs) were compared. The TUs that displayed correlated expression were used to identify putative *cis*-regulatory elements by searching the upstream regions of the TUs for conserved motifs. Predicted motifs were extended and refined by searching for motifs in the upstream regions of additional TUs with correlated expression. In this way, *cis*-acting elements were identified for 41 regulons consisting of at least four TUs (correlation > 0.7). This set of regulons included the known regulons of CtsR and LexA, but also several novel ones encompassing genes with coherent biological functions. Visualization of the regulons and their connections revealed a highly interconnected regulatory network. This network contains several subnetworks that encompass genes of correlated biological function, such as sugar and energy metabolism, nitrogen metabolism and stress response.

Received 30 May, 2010; accepted 31 August, 2010. *For correspondence. Email: michiel.wels@nizo.nl; Tel. (+31) 318 659 674; Fax (+31) 318 650 400.

Introduction

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)] has provided an opportunity to study large regulatory networks of organisms. Complete regulatory networks have been studied in model microbes like *Escherichia coli* and *Saccharomyces cerevisiae* (Hartemink *et al*., 2002; Lee *et al*., 2002; Shen-Orr *et al*., 2002; Bar-Joseph *et al*., 2003; Covert *et al*., 2004; Luscombe *et al*., 2004). These first analyses were mainly based on genomics data, while later efforts included knowledge gathered from databases with curated information on regulatory interactions [e.g. *in vitro* transcription factor (TF)-DNA binding assays] to refine the predicted network (Gutierrez-Rios *et al*., 2003; Herrgard *et al*., 2004; Luscombe *et al*., 2004).

A single microarray dataset provides a snapshot of the complete transcription profile of a cell and therefore it is an extremely valuable source of information for unravelling regulatory networks. However, individual microarray datasets describe only a co-occurring change in the expression of individual genes, which does not automatically imply consistent co-regulation involving a common regulator (e.g. a TF). Correlation analysis of expression and regulation of genes using multiple transcriptome datasets does enable the enrichment of co-regulated genes (Eisen *et al*., 1998).

Another way to obtain insight in the regulatory network of one or more organisms is by *in silico* detection of (conserved) *cis*-acting elements, representing for instance the DNA binding sites of TFs. In this approach, the upstream regions of a group of genes predicted to have the same *cis*-acting element (e.g. on basis of their co-regulation determined by microarray analysis) are analysed using pattern recognition tools such as Gibbs sampling (Thompson *et al*., 2003) or expectation maximization (Bailey and Elkan, 1994). Combining the knowledge of a shared regulatory binding site (*cis*-acting element) with the observed correlated change in the expression of genes allows the identification of the genes

that are co-regulated (Bussemaker *et al*., 2001; Keles *et al*., 2002; Conlon *et al*., 2003). Subsequently, the identified regulatory elements can be used to scan the genome(s) of interest in order to predict the full complement of a regulon. Although these computational 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 transcriptome data originating from different, non-related studies to predict in part the regulatory network of *Lactobacillus plantarum*. Transcriptional units (TUs) (i.e. single gene and multiple gene operons) with correlated expression were identified, and subsequently common *cis*-acting elements within the upstream DNA sequences were determined, starting with the three TUs with highest expression correlation. Then the sets were expanded with additional candidate-regulon members on basis of shared regulatory motifs and (partial) expression correlation.

In this way, a total of 41 sets of co-regulated genes consisting of at least four different TUs were identified. This study shows that correlation analysis of co-regulation in multiple transcriptome datasets combined with *cis*-element prediction provides a valuable strategy for the prediction of regulons and regulatory networks.

Results

Identification of co-regulated TUs and the associated regulatory elements

Gene expression data from 77 different experiments of *L. plantarum* WCFS1 present in the Gene Expression Omnibus (GEO) database were used, including growth of wild-type or mutant strains on different sugar sources, fermentative versus respiratory growth, and environmental stresses, to reconstruct a partial regulatory network (see *Experimental procedures*). The precise procedure to predict the regulatory network of *L. plantarum* is summarized in Fig. 1, and starts with the prediction of TUs and subsequent analysis of their correlated expression. A total of 1735 TUs were predicted in the complete genome sequence of *L. plantarum* WCFS1 using a simple distance criterion (see *Experimental procedures*). Analysis of the available microarray datasets showed that for 523 of these TUs at least one gene showed significant elevated expression in at least one of the experiments (Table 1). For 345 of these TUs, at least two TUs displayed a correlated expression modulation above a set threshold (0.7) (Table 1). Each of these 345 TUs and the two TUs with highest correlated expression were clustered into triplets (see *Experimental procedures*) and then the set was filtered for redundancy, resulting in 286 triplets.

For all 286 TU triplets the DNA sequence upstream of the predicted translation start of the first gene was subjected to MEME analysis (Bailey and Elkan, 1994) to predict conserved *cis*-regulatory elements. Subsequently, a MAST (Bailey and Gribskov, 1998) analysis of the upstream regions on all 1735 TUs was performed. This analysis resulted in the prediction of 62 expanded TU sets, which are triplets that could be enriched with at least one additional TU sharing both a regulatory element and showing expression correlation > 0.7 with the TUs in the original triplet. These TU sets were used, together with the original TUs, to refine the regulatory element sequence (see *Experimental procedures*). Finally, MAST searches were performed with the refined regulatory motifs related to the expanded TU sets to identify all occurrences of the motif (both with correlated as well as non-correlated expression) in the genome. All TUs that showed both high correlation of expression (> 0.7) and shared an upstream motif with a *P*-value $< 1.0e^{-07}$ were regarded as regulon members. The complete set of regulons was checked for redundancy and duplicates were cleared from the set, resulting in a final prediction of 50 different sets of TUs ('regulons'). Nineteen regulons were reduced to less than four TUs during this motif refinement procedure and were not further investigated in this study (see *Experimental procedures*).

A regulon size distribution graph was made for the remaining 31 different regulons (Fig. 2). The majority [23 (75%)] of the predicted regulons has a size between four and six TUs. This observation is in agreement with the commonly accepted notion that only a limited number of so-called global regulators exist in bacteria (Martinez-Antonio and Collado-Vides, 2003). The largest regulon that could be identified on basis of our data in *L. plantarum* consisted of nine different TUs, encompassing a total of 19 genes. A summary of the regulon analysis can be found in Table 1. All identified regulons were stored in a database that can be accessed at http://www.cmbi.ru.nl/regulatory_network/. Several of the predicted regulons harbour genes with a clearly coherent biological role, like stress response, carbohydrate utilization or nitrogen metabolism and represent known regulons in various bacteria.

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

All predicted regulons were visualized in Cytoscape (http://www.cytoscape.org) in order to reconstruct an initial regulatory network of *L. plantarum* WCFS1 (Figs 3 and S1 for detailed view). In the network TUs are displayed as nodes, and regulon members are connected by edges. Analysis of the regulatory network disclosed some remarkable characteristics. First, the overall network can

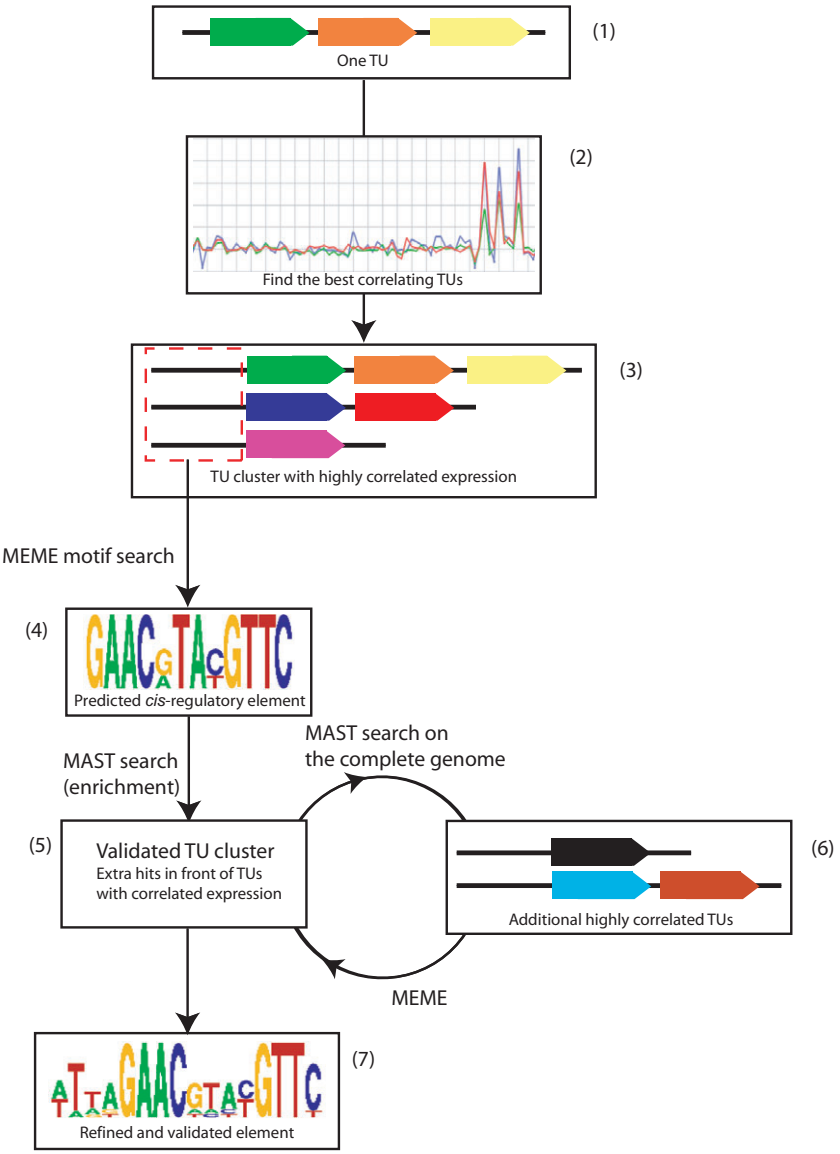


Fig. 1. Flowchart of the followed procedure. 1 = query TU, 2 = identification of co-regulated TUs in a large set of experiments, 3 = selection of upstream sequences of co-regulated TU cluster, 4 = shared *cis*-element detection, 5 = validation of the TU sets on basis of motif occurrence and correlated expression, 6 = iteration procedure to expand or complete regulon, 7 = refined regulon *cis*-element.

be divided into eight smaller subnetworks, some of which consist of a single, isolated regulon. Other regions in the regulatory network form a more highly intertwined and interacting web, sharing interactions between up to 10 regulons in a single connected regulatory (sub)system.

Table 1. Overview of TU and regulon prediction.

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

The connectivity varies in complexity, ranging from regulon pairs that share only a single TU, to pairs of regulons that contain up to seven common TUs. This dense interconnected organization is in line with the dense overlapping regulon structures found in the analysis of the regulatory network of *E. coli* (Shen-Orr *et al.*, 2002). As expected, inspection of the functional annotation of the genes associated with a single regulatory motif showed coherence. Moreover, we observed that several combinations of motifs/regulons displayed functional coherence as well. As an example, TU_1398 (i.e. *lp_3009*, *lp_3010* and *lp_3011*), which encodes a cellobiose PTS and a 6-phospho-beta-glucosidase was found to be part of five different regulons. In fact, many TUs were found to be part of more than a single regulon and hence their upstream regions should contain multiple yet

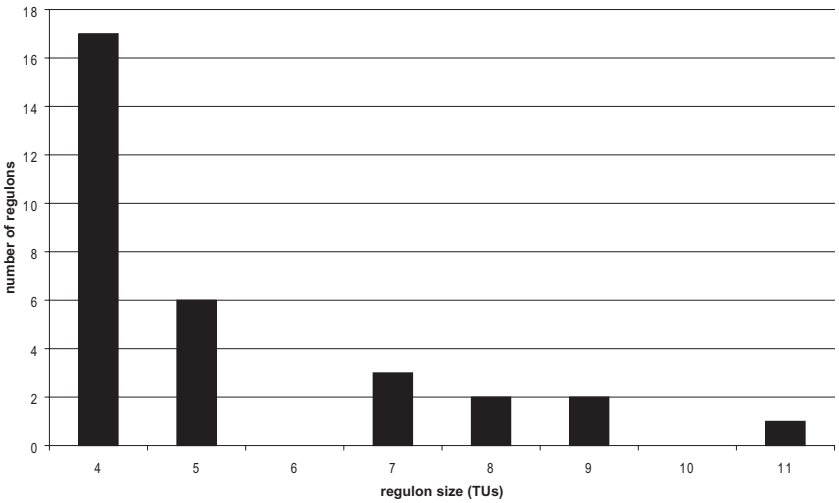


Fig. 2. Regulon size distribution. Regulons of a size < 3 were not regarded (see methods).

different regulatory motifs. The two largest regulatory network subnetworks identified in *L. plantarum* encompass 8 and 10 regulons and display a large degree of functional coherence in terms of functional annotation of the genes contained within the regulatory network subnetwork. In the following paragraphs, several of the regulons will be described in more detail.

CtsR regulon

An example of an isolated regulon is that of CtsR. The regulon contains four stress-related TUs. Three of these TUs (TU_366, TU_601 and TU_884) each consists of a single gene involved in the Clp protease complex (*clpP*, *clpE* and *clpB*, respectively), while the fourth TU contains a single gene annotated as small heat shock protein (*lp_0129*/*hsp1*). The predicted *cis*-element of this regulon is a perfect direct repeat (AAGGTCA-(N3)-AAGGTCA) and strongly resembles the consensus binding site (GGTCAAANANGGTCAAA) of CtsR, as described for *Bacillus subtilis* in Derre and colleagues, 1999. Upstream of TU_884 (i.e. *lp_1903/clpB*) two copies of the corresponding motif were identified (*P*-values $4e^{-07}$ and $5e^{-10}$). CtsR is a stress response regulator known to be involved in Clp activation in several different *Firmicutes*, e.g. *B. subtilis* (Kruger and Hecker, 1998), *Staphylococcus aureus* (Chastanet *et al*., 2003), *Streptococcus pneumoniae* (Chastanet *et al*., 2001) and *Lactococcus lactis* (Varmanen *et al*., 2000). The involvement of CtsR in regulation of this regulon could be confirmed, as 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 members c*lpE* and c*lpB* in *L. plantarum* were shown to be regulated by CtsR in *B. subtilis* (Kelley, 2006a). However, the genome of *B. subtilis* appears to lack a gene encoding an ortholog of the *L. plantarum hsp1* gene, showing that CtsR of *B. subtilis* and *L. plantarum* does not control completely 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* (Kruger and Hecker, 2003). In a recent study it was experimentally proven that regulation of *hsp1* by CtsR occurs in *L. plantarum* WCFS1 (Fiocco *et al*., 2010).

SOS response regulon

The 19 genes of the largest regulon we identified belong to several functional classes, including 'DNA metabolism' (six genes), 'transcription', 'protein synthesis' and 'regulatory functions' (all three categories each one gene), while the residual genes belonged to the category of hypothetical proteins but with putative functions such as segregation helicase (*lp_1543*) and exopolyphosphatase-related protein (*lp_2279*). The regulon motif identified was a highly conserved palindromic sequence (GAAC-(N4)-GTTC), resembling the binding site of LexA (or DinR) involved in the regulation of the SOS response in various organisms [for a good review on the SOS response, see (Kelley, 2006b)]. Notably, the gene for LexA itself (*lp_2063*) also appears 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* (Fernandez De Henestrosa *et al*., 2000; Au *et al*., 2005). The regulatory process of DNA damage repair is based upon cleavage of LexA by RecA, a protein activated by binding to single-stranded DNA. Analysis of the 91 bp intergenic region between *recA* and *cinA* showed that there is a conserved LexA binding site in the upstream sequence of *recA*, connecting this gene to the SOS locus. *RecA* being a member of the SOS regulon is in line with the regulon organization

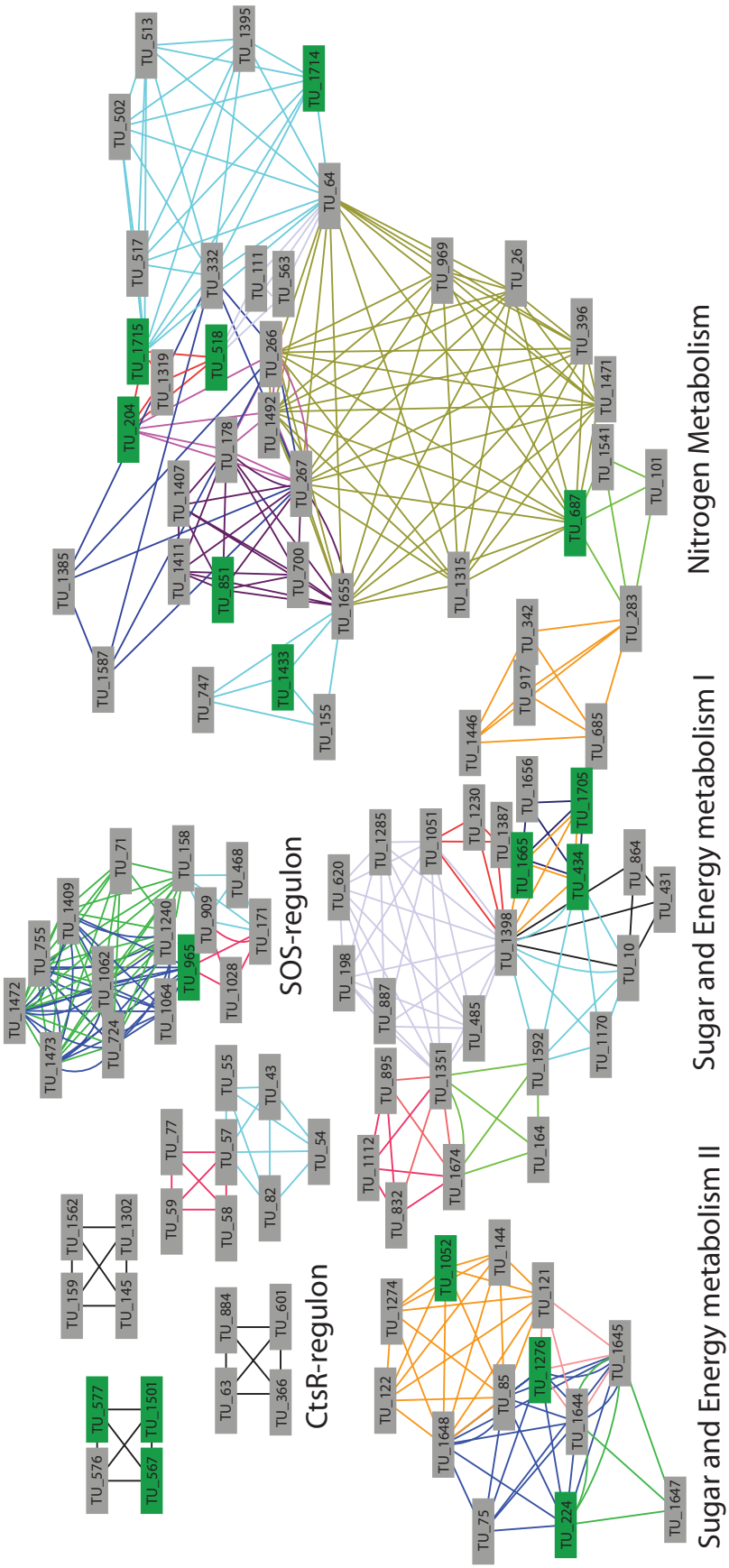


Fig. 3. Regulatory network of *L. plantarum*. Regulons displayed using Cytoscape (http://www.cytoscape.org). Nodes (rectangles) represent different TUs; edges connect TUs that are member of the same regulon. Connections originating from the same regulon share the same color. Nodes describing a TU that encoded at least one regulatory protein were colored green. The supplementary information contains a sif file (for use within cytoscape) and a high quality figure of the network (Figure S1).

described in other bacteria (Kelley, 2006). At first, we did not find *recA* to be part of the SOS regulon in *L. plantarum* because it was grouped into one TU with the upstream located gene (*cinA*). Manual inspection of the expression correlation between *recA* and *cinA* showed that these genes displayed a low correlation in expression (0.50) suggesting that these two genes are probably not part of the same TU. Detection of a LexA binding site between the two genes and the observation that *recA* displayed a highly correlated expression with several other members of the SOS regulon (e.g. *umuC*: 0.82; *lexA*: 0.75; *dinP*: 0.88) supports the hypothesis that *recA* and *cinA* are not members of the same TU and that *recA* is in fact a true member of the SOS regulon in *L. plantarum*.

In addition to the 11 *cis*-acting elements in the upstream regions of SOS regulon TUs with a highly correlated expression, 18 additional hits were found to this motif in the genome-wide MAST search. In these 18 cases, the downstream located TUs had a correlated expression below 0.7 with the predicted SOS regulon. Three 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 2).

The *L. plantarum* regulon composition was compared with the SOS regulon in *B. subtilis* [obtained from the DBTBS (Makita *et al*., 2004)] by pairwise BLAST analyses of the two (genome-based) proteomes. 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 for only 11 there appeared to be orthologs in *B. subtilis*. Moreover, only four of these genes have been reported to belong to the LexA regulon in *B. subtilis*. Of the additional putative regulon members with a LexA binding site in *L. plantarum* (Table 2, correlation between 0.5 and 0.7), five orthologs were identified in *B. subtilis*, of which four are subject to LexA regulation in that host [i.e. *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 (Kelley, 2006).

Novel regulons

The largest subnetwork of 10 regulons contains five TUs that are highly connected (present in three 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 arginine/glutamate biosynthesis while TU_1655 (lp_3497–lp_3499) encodes genes responsible for synthesis of aromatic amino acids. The two residual TUs contain genes for less clear functions, as they encode a two-component regulator (lp_0130–lp_0131) and a transporter with unknown specificity (lp_3183). Interestingly, upstream of this last gene a TU is found that contains two genes both encoding a branched chain amino acid transporter (*lp_3184* and *lp_3185*). The strong connection between these five TUs suggests that they may all play a role in amino acid biosynthesis or uptake.

The second largest regulatory subnetwork contains eight connected regulons, encompassing 25 TUs containing a total of 61 genes (Table 3 and http://www.cmbi.ru.nl/regulatory_network/). Many of these genes encode sugar metabolism-related functions, including eight genes encoding different PTS subunits, three genes encoding sugar metabolism-related regulatory proteins and 10 polysaccharide or sugar metabolism-related enzymes. In addition, this regulatory subnetwork includes 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 3). Next to these, additional genes were encompassed in this subnetwork that could potentially be related to sugar metabolism, i.e. three transporters with unknown substrate specificity and four genes encoding cell envelope proteins (*lp_2795*, *lp_2796*, *lp_1303a*, *lp_2921*). Sixteen genes encode hypothetical proteins of unknown function and 17 genes were clearly not involved in sugar metabolism based on their annotation.

Analysis of the predicted *cis*-acting elements within this sugar metabolism regulatory network revealed that there is only a limited overlap in regulatory 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). However, differences were observed between the scoring matrices. These differences generate different MAST results, leading to the prediction of two different regulons. Iteration of the motif detection did not result in a merger of the two regulons. The identified 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 the consensus sequence is also representative for the binding site of different members of the LacI family of regulators in *L. plantarum* (Francke *et al*., 2008). The regulators of this family are known to be involved in the regulation of many different sugar metabolism genes. In addition to the two *cre*-like occurrences, only one addi-

Table 2. SOS regulon of *L. plantarum*.

TU	Gene	Function	Main class	*P*-value	Gene expression correlation
True regulon members (correlation > 0.70)					
1473[a]	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[a]	lp_3141	Unknown	Hypothetical proteins	2.1e-11	0.85
724	lp_1543 (*cshA2*)	One 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[a]	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[a]	lp_2063 (*lexA*)	transcription repressor of the SOS regulon	Regulatory functions	4.2e-08	0.79
1064[a]	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[a]	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[a]	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		
No transcriptome data available					
469	lp_0961	Unknown	Hypothetical proteins	2.7e-10	n/a
1065[a]	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[a]	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[a]	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

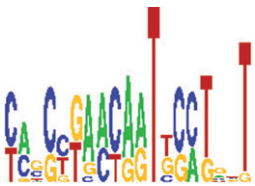
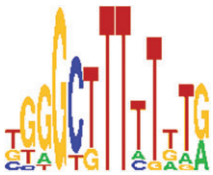
a. Hit found in the intergenic region between two divergently transcribed TUs, possible false-positive.
All regulon members are listed (TUs and encoded genes). *P*-values of MAST hits (motifs) in upstream regions are shown, together with correlation values of gene expression.
N/A, no expression data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

tional case of overlap was observed between the predicted motifs. Two different motifs were found to partially overlap in the upstream region of TU_1674. This was not immediately apparent from the motif logos as the motifs were identified on two different strands. Next to TU_1674, we found that TU_1351 also shares these two *cis*-acting elements; but in this case 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 (i.e. different

Table 3. Regulons clustered in the sugar metabolism regulatory network subunit.

Motif	TUs	Main classes
	TU_832	Energy metabolism (glycolysis)
	TU_895	Transport
	TU_1112	Cell envelope (lipoproteins)/transport
	TU_1351	Hypothetical proteins
	TU_1674	Central intermediary metabolism
	TU_164	Hypothetical proteins
	TU_1351	Hypothetical proteins
	TU_1592	Other categories (prophage-related)
	TU_1674	Central intermediary metabolism
	TU_198	Hypothetical proteins
	TU_485	Hypothetical proteins
	TU_620	Cell envelope (cell surface)
	TU_887	Transport (multidrug)
	TU_1285	Cell envelope (LPXTG)
	TU_1051	Hypothetical proteins
	TU_1351	Hypothetical proteins
	TU_1398	Transport (PTS)/energy metabolism (sugars)
	TU_1170	Amino acid biosynthesis (histidine)
	TU_10	Central intermediary metabolism (polysaccharides)
	TU_434	Regulatory functions (BglB)/transport (PTS)/energy metabolism (sugars)
	TU_1398	Transport (PTS)/energy metabolism (sugars)
	TU_1592	Other categories (prophage-related)
	TU_431	Cellular processes (chaperones)
	TU_864	Hypothetical proteins
	TU_10	Central intermediary metabolism (polysaccharides)
	TU_1398	Transport (PTS)/energy metabolism (sugars)
 [a]	**TU_434**	Regulatory functions (BglB)/transport (PTS)/energy metabolism (sugars)
	TU_1398	Transport (PTS)/energy metabolism (sugars)
	TU_1665	Regulatory functions/transport (PTS)/hypothetical proteins
	TU_1705	Regulatory functions/transport (PTS)/energy metabolism (sugars and general)
 [a]	TU_1656	Energy metabolism/hypothetical proteins
	TU_434	Regulatory functions (BglB)/transport (PTS)/energy metabolism (sugars)
	TU_1665	Regulatory functions/transport (PTS)/hypothetical proteins
	TU_1705	Regulatory functions/transport (PTS)/energy metabolism (sugars and general)
	TU_1230	Transport
	TU_1387	Cell envelope (teichoic acid biosynthesis)
	TU_1051	Hypothetical proteins
	TU_1398	Transport (PTS)/energy metabolism (sugars)

a. LacI family motifs.
TUs in bold are shared among multiple different regulons. More details on these proposed regulons can be found at www.cmbi.ru.nl/regulatory_network.

sugar sources, environmental stresses like H_2O_2 and fermentative vs. respiratory growth) and mutant derivative strains showed to be of great value for obtaining data-driven insight in the regulatory network of *L. plantarum*. Correlation in expression data could be used to identify co-regulated TUs. Combined with detection of shared *cis*-acting regulatory elements, a database of predicted regulons was constructed that allowed reconstruction of a partial regulatory network of *L. plantarum* WCFS1. In many cases the *cis*-acting elements were not previously identified. Moreover, functional coherence within regulons and regulatory subnetworks became apparent, supporting the biological relevance of the identified network.

The two largest subsystems within the partial regulatory network seem to represent carbon (sugars) and nitrogen (amino acid) metabolism. CodY, the master regulator in nitrogen regulation in many *Firmicutes* was found to be absent in *L. plantarum* (Martinez-Antonio and Collado-Vides, 2003). It is interesting to see if alternative candidate regulators are present in this nitrogen metabolism subnetwork. Analysis of the subnetwork revealed the presence of eight genes encoding regulatory functions (*lp_0396*, *lp_0889*, *lp_1092*, *lp_1443*, *lp_1821*, *lp_3079*, *lp_3649* and *lp_3650*).

Although this study shows that correlated expression over a large set of microarray data can be of great help in unravelling a partial regulatory network of an organism, the quality, amount and source of microarray data probably have a great influence on the usefulness of the data. Although this analysis incorporated the data of more then 70 microarrays, only 802 genes (< 25%) displayed a sufficient level of differential expression in any of the experiments, to be included in the co-regulation analysis performed here. This relatively low number of differentially expressed genes is probably the result of a relatively low variability within the experimental conditions related to the transcriptome dataset. Many of the experiments included in the dataset were performed on comparable media and cells were sampled in the same growth phase. An increase in the variability of the experimental conditions should result in more variability in the expression pattern of individual genes. This will result in the incorporation of a higher number of genes in the initial analysis and thus increase the scope (size) and resolution of the network. Eventually, an increase in the number of regulons could potentially link the different subnetworks 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 (i.e. 41). As our method is fully automatic, the regulatory network can be easily fine-tuned and updated when the amount of transcriptomics data increases.

Surprisingly, the network described in this study lacks the identification of big (global) regulons. it seems unlikely that *L. plantarum* does not contain any global regulons, as global regulators (and thus global regulons) appear to be present in (almost) all bacteria (Martinez-Antonio and Collado-Vides, 2003). The lack of detection of these regulons could be due to the fact that global regulators often regulate the expression of genes that are also regulated by locally acting regulators (Chauvaux *et al*., 1998; Francke *et al*., 2008). The regulatory effect of the local regulators will disturb the signal (and thus the correlation of expression) of the global regulator, making it impossible to detect these regulons. It was recently suggested that this phenomenon could very well occur between CcpA and many other LacI family members in *L. plantarum* (Francke *et al*., 2008).

The partial network displays a degree of connectivity that is comparable with earlier observations in other bacteria. Additional, more detailed information on TF-*cis* regulatory element interactions could enable us to dissect this interconnected network to the 'network motifs' of regulation as suggested by Shen-Orr and colleagues (Shen-Orr, *et al*., 2002). In some cases the TF binding to a *cis*-acting element was predicted on basis of literature data (e.g. SOS response, regulation of Clp proteases). In other regulons the lack of literature studies did not allow linking a *cis*-acting element to a TF. Other strategies, like analysing the gene neighbourhood of the TUs in a regulon or performing ChIP-chip experiments to obtain TF binding data could lead to the prediction of a regulatory network with a higher resolution. The regulatory network created in this study can be of great help in the analysis of a transcriptomic response by displaying the data on the reconstructed network. Moreover, combining the knowledge in this regulatory network and connecting it to the reconstructed metabolic network of *L. plantarum* (Teusink *et al*., 2006) will help to increase our understanding of the global gene expression and metabolic adaptation of *L. plantarum* to changes in its environment. The reconstruction of the regulatory network will be of great help to elucidate the processes that underlie specific *in situ* behaviour, for example, during food fermentation processes or gastro-intestinal tract residence. Moreover, the combined reconstructed networks could be used to rationalize the discovery of targets for optimizing culture performance and for improving strain robustness.

Experimental procedures

Prediction of TUs

Transcription units were predicted with the distance-based method described by Wels and colleagues, 2006, except that the intergenic distance within a TU was expanded from

< 50 nt to < 100 nt to decrease the likelihood of splitting one functional TU into multiple, smaller TUs. Although this increase in intergenic distance will decrease the number of TUs with correlated expression, TUs that are inappropriately divided into multiple smaller TUs would 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. The power of using solely intergenic distance as a determinant for membership of a single TU was demonstrated by the comprehensive analysis of operon prediction methods (Brouwer *et al.*, 2008).

Expression data

Expression data were obtained from the GEO (Edgar *et al.*, 2002) at 1 January 2010. This set contained 77 microarray experiments of *L. plantarum* WCFS1, using Agilent oligo-based arrays. Other (previously designed) microarray platforms were not regarded in this study because of the significant reduction in quality. The tested experimental conditions were highly variable and ranged from stress conditions (such as ethanol and peroxide challenge) to knockout or overexpression of specific (metabolic) genes. The GEO accession codes of the used datasets are GSM136883–136888 [GSE5882 (Saulnier *et al.*, 2007)], GSM206844–GSM206852 [GSE8348 (Serrano *et al.*, 2007)], GSM215123–GSM215128 (GSE8672), GSM217127–GSM217132 (GSE8743), GSM445770–GSM445781 (GSE17847), GSM 457827–GSM457837 (GSE18339), GSM457838–GSM457842 (GSE18340), GSM458130–GSM458135 (GSE18354), GSM459362–GSM459371 (GSE18432), GSM459519–GSM459534 (GSE18435). Some other series of microarray data found in GEO were not regarded as a result of experimental overlap (GSE8744, too much overlap with GSE8743) or the use of other strains (GSE18435, performed with both *L. plantarum* WCFS1 but also *L. plantarum* NC8). 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).

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 every TU for which gene expression data were available, the two best correlating TUs were extracted from the database and grouped into one 'triplet' of TUs. 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 and optimization

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. MEME software (Bailey and Elkan, 1994) was used to predict regulatory motifs from upstream regions of 300 nt preceding the translation start of the first genes within the selected TUs. MEME default settings were applied except for: minimum length 5 nt, maximum length 20 nt, four different motifs, found in at least two out of three sequences and find motifs on both strands (-revcomp). MAST (Bailey and Gribskov, 1998) was used to search given MEME output motifs in the upstream regions of all predicted TUs. If additional motifs were detected in the upstream sequence of other TUs, these upstream regions were used, in combination with those of the original triplet, to generate a refined MEME motif and this refined motif was then used to search the set of TUs again with MAST. This procedure was iterated until no additional TUs sharing the upstream element and with significant correlation was recovered. Although the procedure of identifying regulons by iteratively searching for TUs with a shared motif and a highly correlated expression was focused on finding additional members associated with the original triplets and thus generate regulons of at least four TU members, 19 regulons were identified of a size smaller than four TUs. The results of the MEME analysis for these 19 regulons appeared to consist of motifs that scored worse than the final, more stringent MAST cut-off that was applied. These regulons were therefore not further investigated in this study.

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 subclasses 'Not conserved: other', 'Conserved: other', 'Conserved: putative function' within the main class 'Hypothetical proteins' and the subclass 'Unknown substrate' within the main class 'Transport and Binding' were not considered relevant, as 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 recovered regulons.

Cytoscape visualization

Visualization was performed by loading tab-delimited text files into the Cytoscape software package (Shannon *et al.*, 2003). Two TUs were connected by an edge if they shared both expression (correlation > 0.7) and a motif (*P*-value < $1.0e^{-07}$). Data were first ordered using the spring embedded sorting algorithm in the tool. The network was structured into smaller units by manual inspection and alteration. Colouring of the edges (based on shared regulon) and nodes (if the TU consisted of at least one regulatory protein) was performed manually.

References

Au, N., Kuester-Schoeck, E., Mandava, V., Bothwell, L.E., Canny, S.P., Chachu, K., *et al.* (2005) Genetic composition

of the *Bacillus subtilis* SOS system. *J Bacteriol* **187:** 7655–7666.

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.

Bailey, T.L., and Gribskov, M. (1998) Combining evidence using *P*-values: application to sequence homology searches. *Bioinformatics* **14:** 48–54.

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21:** 1337–1342.

Brouwer, R.W., Kuipers, O.P., and van Hijum, S.A. (2008) The relative value of operon predictions. *Brief Bioinform* **9:** 367–375.

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

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.

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.

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.

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 USA* **100:** 3339–3344.

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.

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.

Edgar, R., Domrachev, M., and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30:** 207–210.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **95:** 14863–14868.

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.

Fiocco, D., Capozzi, V., Collins, M., Gallone, A., Hols, P., Guzzo, J., *et al*. (2010) Characterization of the CtsR stress response regulon in *Lactobacillus plantarum*. *J Bacteriol* **192:** 896–900.

Francke, C., Kerkhoven, R., Wels, M., and Siezen, R.J. (2008) A generic approach to identify Transcription Factor-specific operator motifs; Inferences for LacI-family mediated regulation in *Lactobacillus plantarum* WCFS1. *BMC Genomics* **9:** 145.

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.

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* **2002:** 437–449.

Herrgard, M.J., Covert, M.W., and Palsson, B.O. (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol* **15:** 70–77.

Keles, S., van der Laan, M., and Eisen, M.B. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics* **18:** 1167–1175.

Kelley, W.L. (2006a) Lex marks the CtsR heat shock regulon of *Bacillus subtilis*. *Mol Microbiol* **62:** 581–593.

Kelley, W.L. (2006b) Lex marks the spot: the virulent side of SOS and a closer look at the LexA regulon. *Mol Microbiol* **62:** 1228–1238.

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.

Kruger, E., and Hecker, M. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *J Bacteriol* **6:** 482–489.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804.

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.

Makita, Y., Nakao, M., Ogasawara, N., and Nakai, K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res* **32:** D75–D77.

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

Saulnier, D.M., Molenaar, D., de Vos, W.M., Gibson, G.R., and Kolida, S. (2007) Identification of prebiotic fructooligosaccharide metabolism in *Lactobacillus plantarum* WCFS1 through microarrays. *Appl Environ Microbiol* **73:** 1753–1765.

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.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504.

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.

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.
Thompson, W., Rouchka, E.C., and Lawrence, C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* **31:** 3580–3585.
Varmanen, P., Ingmer, H., and Vogensen, F.K. (2000) ctsR of *Lactococcus lactis* encodes a negative regulator of *clp* gene expression. *Microbiology* **146** (Part 6)**:** 1447–1455.
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.

Supporting information

Additional Supporting information may be found in the online version of this article:

Fig. S1. Regulatory network of *L. plantarum*. High quality figure of figure 3. Regulons displayed using Cytoscape (http://www.cytoscape.org). Nodes (rectangles) represent different TUs; edges connect TUs that are member of the same regulon. Connections originating from the same regulon share the same color. Nodes describing a TU that encoded at least one regulatory protein were colored green.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

