Large Intergenic Cruciform-Like Supermotifs in the *Lactobacillus plantarum* Genome

Michiel Wels,1,2,3* Roger S. Bongers,1,3 Jos Boekhorst,2 Douwe Molenaar,1,3 Mark Sturme,4 Willem M. de Vos,1,4 Roland J. Siezen,1,2,3 and Michiel Kleerebezem1,3,4

TI Food and Nutrition, Wageningen, The Netherlands1; Radboud University Nijmegen-Medical Centre/NCMILS, Nijmegen, The Netherlands2; NIZO Food Research, Ede, The Netherlands3; and Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands4

Received 28 November 2008/Accepted 5 March 2009

Twenty-four *Lactobacillus plantarum* supermotifs (LPSMs) with lengths from ~800 to 1,000 nucleotides were identified in the *L. plantarum* genome. LPSMs were conserved in other *L. plantarum* strains but not in other species. Secondary structure analysis predicted that LPSMs may fold into cruciform-like structures. Preliminary experiments indicate that the LPSMs are transcribed.

Intergenic regions of bacterial genomes are known to contain many small elements, such as promoters, transcription or translation regulation sequences, or terminators, that play a role in gene expression. Occasionally, much larger functional noncoding regions are encountered that contain larger functional genetic elements. These elements are involved in transcription regulation (e.g., riboswitches and T-box elements [8, 20]) or function as trans-acting RNA elements, blocking transcription or translation by binding to their target (e.g., RNAIII in *Staphylococcus aureus* [17, 18]). Another example of well-described noncoding elements is CRISPRs, RNA elements that protect the cell against phage attacks. CRISPRs are known to be distributed in a subset of lactic acid bacteria genomes (3, 10). For these elements, it was shown that they are transcribed to RNA and are processed by the associated enzyme complex named CAS to generate the functional small RNAs that interfere with phage replication or function (4). In addition, there are many examples of repeated elements in bacterial chromosomes with an undefined function (15, 16, 21). The largest repeated elements described to date have been identified exclusively in *Staphylococci* species and have been designated *Staphylococcus aureus* repeat (STAR) elements (5). STAR elements are internally repeated DNA sequences with lengths, in general, of between 300 and 600 nucleotides (nt), which have been found up to 21 times in one genome (5). The function of these STAR elements has not been described in the literature.

In this study, we identify conserved genetic elements in the intergenic regions of *Lactobacillus plantarum*. Initially, all 74 noncoding regions larger than 700 nt were scanned for conserved sequence motifs using MEME software (1) with the settings “any” (any number of repetitions), a minimum number of occurrences of 20, a width range of between 10 and 100 nt, and a maximum of 20 different motifs. This analysis revealed 20 different motifs with various lengths, from 19 to 100 nt. Remarkably, 19 motifs (all smaller than 50 nt) were found to colocalize in the same intergenic loci, indicating the existence of a large conserved intergenic sequence in *L. plantarum*. Detailed manual inspection of the conserved order of these motifs revealed a >800-nt supermotif, which is termed here *Lactobacillus plantarum* supermotif (LPSM).

A MAST analysis (2) on the total chromosomal sequence revealed 26 intergenic regions that encompassed at least 10 consecutive motifs (<50 nt between two consecutive motifs). A hidden Markov model (HMM) was built from the complete alignment of these 26 regions using Muscle software (7) and used to search the complete genomic sequence of *L. plantarum* WCFS1 (using HMMer [6]). In an iterative procedure, hits below the threshold (E value of 0.1) were used to build a refined HMM until no additional hits were found. In total, 66 LPSM hits were identified. All minus-strand hits exactly colocalized with a plus-strand hit, indicating that there are 33 copies of a conserved bidirectional LPSM within the chromosome of *L. plantarum* WCFS1 (for an alignment of the 33 copies, see file S1 in the supplemental material). The LPSM sequences did not coincide with statistical features on the genome, such as GC%, codon adaptation index, or comparative genome hybridization data (gathered from reference 14) (Fig. 1). However, the LPSMs were found to be statistically overrepresented on the second replichore of the genome (chi-square test; *P < 0.05*). 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 are summarized (see Table S1 in the supplemental material). The genomic context and flanking gene order of all LPSMs was compared with those of other related bacteria. Seventeen out of 33 LPSMs were found in regions that share synteny with other genomes (some examples are displayed in Fig. 2), in particular with genomes of other lactic acid bacteria (13). The corresponding intergenic regions in *L. plantarum* were consistently much longer, indicating that LPSMs are uniquely present in these loci in the *L. plantarum* genome. In addition, no hits above the threshold (0.1) were found by HMM searches against all publicly available genomes found in...
the NCBI repository (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/), indicating that the full-length LPSM is unique for *L. plantarum*.

A PCR approach was used to analyze conservation of 15 of the LPSMs among 10 additional *L. plantarum* strains (see file S2 in the supplemental material). PCR amplifications were carried out with an automated thermal cycler (Perkin-Elmer, Shelton, CT) using *Taq* DNA polymerase (Gibco-BRL Life Technologies, Breda, The Netherlands). Primers were designed to specifically anneal to genes that flank 15 of the LPSMs present in the *L. plantarum* WCFS1 genome and were used to amplify the corresponding regions in the chromosomal DNA isolated from 10 other *L. plantarum* strains. Selection of these strains was based on their relative genetic distance to the sequenced strain WCFS1, which was previously determined by gene presence profiling using comparative genome hybridization (14). In 117 cases (out of the total of 150), these PCRs generated detectable amplification products, of which 111 were of a size comparable to the product obtained using *L. plantarum* WCFS1 chromosomal DNA as a template. Of the remaining six products, three appeared to be significantly shorter than the WCFS1 product and three were longer (at least a 500-bp difference). Moreover, none of the LPSMs was found to be unique for WCFS1 only, while all *L. plantarum* strains analyzed contained at least eight LPSMs. Therefore, it can be concluded that LPSMs are generally found on the chromosomes of different *L. plantarum* strains and that their chromosomal locations are largely conserved among strains.

Eight LPSM amplicons obtained from these strains were cloned in *Escherichia coli* using the pGEM-T easy vector system (Promega Corp., Madison, WI), and transformants were readily obtained, suggesting that these LPSM sequences can be stably cloned in this host organism. Sequencing of the inserts was performed (BaseClear, Leiden, The Netherlands) using...
The sequences obtained were aligned with the corresponding *L. plantarum* WCFS1 sequences using Muscle software (7). These sequences displayed a high level (>88%) of sequence identity with the corresponding region found in strain WCFS1. In addition, a maximum likelihood phylogenetic tree (Phyml [9]) of the sequenced LPSMs in combination with the 33 identified *L. plantarum* WCFS1 LPSMs showed clustering for seven out of eight of the sequenced LPSMs with their locus-specific counterparts in *L. plantarum* WCFS1 (see Fig. S1 in the supplemental material), suggesting that the motifs were present in the last common ancestor of the *L. plantarum* strains tested.

The putative secondary structures of the LPSM DNA regions were predicted using the Mfold program (22). The folded structures had free energy values (∆G) ranging from −93 kJ·mol⁻¹ to −149 kJ·mol⁻¹ and an average value of −132 kJ·mol⁻¹ compared to −73 kJ·mol⁻¹ for scrambled LPSM sequences. Mfold predicted similar secondary structures for 24 out of 33 LPSM sequences, structures that were not predicted from the scrambled sequences. The structures were compared manually and divided into different subfamilies. Nineteen LPSMs displayed conserved and typical secondary structures (Fig. 3A). Notably, these structures share remarkable similarity with DNA structures of eukaryotic origin with unknown functions called cruciforms (12). Of the remaining five LPSM sequences, two (LPSM 5 and LSPM 15) appeared to contain an additional hairpin in the "head" region of the LPSM structure, while the structure shown in panel C lacks the "arms" of the LPSM structure.
additional hairpin loop in the top of the cruciform structure (Fig. 3B), while three others (LPSM 14, 20, and 33) were characterized by the lack of “arms” (Fig. 3C). In all cases, these aberrations of the general LPSM structure were the result of deletion or insertion in part of the sequence and are clearly apparent from the multiple alignment of all LPSMs (see file S1 in the supplemental material). The predicted common fold of the LPSM sequence elements provided an alternative structure-based method rather than a sequence-based method for searching similar LPSM structures in other bacterial genomes. Therefore, all intergenic regions of lengths larger than 500 nt were subjected to Mfold structure prediction. Nucleic Acids Res. 31:5131–5136.


