Comparative analysis of proteins with a mucus-binding domain found exclusively in lactic acid bacteria

Jos Boekhorst,[1] Quinta Helmer,[1] Michiel Kleerebezem[2,3] and Roland J. Siezen[1,2,3]

[1]Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen, 6525ED, Nijmegen, The Netherlands

[2]Wageningen Centre for Food Sciences, Wageningen, The Netherlands

[3]NIZO food research, Ede, The Netherlands

Correspondence
Jos Boekhorst
J.Boekhorst@cmbi.ru.nl

Received 8 August 2005
Revised 21 October 2005
Accepted 21 October 2005

Lactic acid bacteria (LAB) are frequently encountered inhabitants of the human intestinal tract. A protective layer of mucus covers the epithelial cells of the intestine, offering an attachment site for these bacteria. In this study bioinformatics tools were used to identify and characterize proteins containing one type of mucus-binding domain, called MUB, that is postulated to play an important role in the adherence of LAB to this mucus layer. By searching in all protein databases 48 proteins containing at least one of these MUB domains in nine LAB species were identified. These MUB domains varied in size, ranging from approximately 100 to more than 200 residues per domain. Complete MUB domains were found exclusively in LAB. The number of MUB domains present in a single protein varied from 1 to 15. In some cases, orthologous proteins in closely related species contained a different number of domains, indicating that repeats of the domain undergo rapid duplication and deletion. Proteins containing the MUB domain were often encoded by gene clusters that encode multiple extracellular proteins. In addition to one or more copies of the MUB domain, many of these proteins contained other domains that are predicted to be involved in binding to and degradation of extracellular components. These findings strongly suggest that the MUB domain is an LAB-specific functional unit that performs its task in various domain contexts and could fulfil an important role in host–microbe interactions in the gastrointestinal tract.

INTRODUCTION

The human gastrointestinal tract is home to at least 400 different bacterial species (Eckburg *et al.*, 2005; Servin, 2004). A protective layer of mucus, consisting of a complex mixture of large, highly glycosylated proteins (mucins) (Dekker *et al.*, 2002) and glycolipids, covers the epithelial cells of the intestine and offers an attachment site for the bacteria colonizing the intestine. These bacteria play an important role in maintaining normal gut functionality and in the resistance of the host against pathogenic micro-organisms (Hooper & Gordon, 2001), and some may use mucins as their major carbon and energy source (Aryanta *et al.*, 1991; Bayliss & Houston, 1984; Sonnenburg *et al.*, 2005). Certain strains of LAB may promote health in man and animals (Reid *et al.*, 2003), and many have been shown to adhere to intestinal mucus (Servin, 2004). In most cases, this adhesion has been shown to be mediated by proteins (Coconnier *et al.*, 1992; Conway & Kjelleberg, 1989; Roos & Jonsson, 2002).

An extracellular mucus-binding protein of *Lactobacillus reuteri* 1063 was identified by Roos & Jonsson (2002). This protein contains two different types of repeats of approximately 200 aa, present in eight and six copies, shown to be responsible for the adherence to intestinal mucus. More recently, Pretzer *et al.* (2005) have identified a protein of *Lactobacillus plantarum* WCFS1 that contains a domain similar to the mucus-binding (MUB) domains identified by Roos & Jonsson (2002) and is involved in the adherence to mannose, which is a constituent of mucin glycosylation moieties. This domain is partly similar to the MucBP domain from the Pfam database (Bateman *et al.*, 2004), but is significantly different in size, sequence and phylogenetic distribution.

The mucus-binding proteins of both *Lb. plantarum* and *Lb. reuteri* have characteristics typical of cell-surface proteins of Gram-positive bacteria: an N-terminal signal peptide

Abbreviations: HMM, hidden Markov model; LAB, lactic acid bacteria; MUB, mucus-binding (domain).

Supplementary tables and figures are available with the online version of this paper.

targeting the protein for secretion and a C-terminal sortase recognition site targeting the protein for covalent attachment to the peptidoglycan layer at the outside of the bacterial cell (Ton-That *et al.*, 2004).

The modification and recombination of existing functional modules plays an important role in evolution of protein function (Doolittle & Bork, 1993). Extracellular proteins are often large proteins consisting of many of these modules or domains (Bork, 1991). The identification and characterization of these domains can play an important role in elucidating the function of extracellular proteins. We have searched bacterial genome sequences and the UniProt protein database for potential mucus-binding proteins based on the sequence of the MUB domains of *Lb. reuteri* and *Lb. plantarum*. We discuss the properties and variability of the MUB domain and the putative role of the domain as a functional unit.

METHODS

Sequence information. Sequence information was obtained from the NCBI bacterial genome database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome), the universal protein resource UniProt (Bairoch *et al.*, 2005) and the ERGO database (Overbeek *et al.*, 2003). From the ERGO database we used genome information for *Enterococcus faecium* DO, *Lactobacillus brevis* ATCC 367, *Lactobacillus casei* ATCC 33323, *Lactobacillus delbrueckii* subsp. *bulgaricus* ATCCBAA-356, *Lactococcus lactis* subsp. *cremoris* SK11, *Leuconostoc mesenteroides* ATCC 8293, *Oenococcus oeni* PSU-1 and *Pediococcus pentosaceus* ATCC 25745. A list of the species present in the NCBI bacterial genome database at the time of this analysis can be found in supplementary Table A (available with the online version of this paper).

Sequence analysis. Sequence similarity was detected with BLAST (Altschul *et al.*, 1990) while multiple sequence alignments were made with Muscle (Edgar, 2004). The HMMER package (Durbin *et al.*, 1998) was used to construct Hidden Markov models (HMMs) of MUB domains and to scan protein sequences with HMMs. HMMs from the Pfam (Bateman *et al.*, 2004), SMART (Letunic *et al.*, 2004) and Superfam (Gough *et al.*, 2001) databases were used to identify other known domains in proteins with identified MUB domains. HMMs were compared with HHsearch (Soding, 2005).

Sortase recognition sites with LPxTG-type motifs were predicted with a recently developed HMM (Boekhorst *et al.*, 2005). Conserved sequence patterns were identified with MEME and MAST (Bailey & Elkan, 1994). Proline-rich amino acid stretches were identified using simple Python scripts. Protein secondary structure predictions were done with PsiPred (McGuffin *et al.*, 2000). The EMBOSS package (Rice *et al.*, 2000) was used to scan for repetitive DNA sequences. DNA secondary structure predictions were done with MFOLD (Santa Lucia, 1998).

Identifying putative mucus-binding proteins. An initial set of potential mucus-binding proteins was identified by searching amino acid sequences obtained from the sources mentioned above with an HMM based on the MUB domains of protein Mub of *Lb. reuteri* (Roos & Jonsson, 2002) and an HMM based on the domains of protein lp_1229 of *Lb. plantarum* WCFS1 (Kleerebezem *et al.*, 2003; Pretzer *et al.*, 2005). Hits with an e-score of 1e-10 or lower were considered putative MUB-domain proteins. The amino acid sequences of these proteins were scanned for conserved protein motifs with MEME. The exact position of the MUB domains in the identified proteins was determined based on the results of the MEME and HMM analyses, and on multiple sequence alignments of highly similar MUB-domain-containing proteins. Multiple sequence alignments of the individual MUB domains were used to create new HMMs, which were subsequently used to search for additional MUB-domain proteins. This iterative process was repeated until no additional MUB-domain proteins were detected in the bacterial genomes or the UniProt database.

RESULTS AND DISCUSSION

Defining the boundaries of the MUB domain

The putative MUB domain of the protein lp_1229 of *Lb. plantarum* WCFS1 consists of approximately 100 aa, while the MUB domain of protein Mub of *Lb. reuteri* is almost 200 residues in length (Kleerebezem *et al.*, 2003; Pretzer *et al.*, 2005; Roos & Jonsson, 2002). This difference in size implies a discrepancy in the definition of the domain boundaries in the *Lb. plantarum* or *Lb. reuteri* mucus-binding proteins. To create domains of a uniform size would require either merging of every second repeat of the *Lb. plantarum* mucus-binding protein with its neighbour or splitting the *Lb. reuteri* domain in two. However, our sequence analysis suggests that the mucus-binding building blocks of the mucus-binding proteins do in fact vary in size. A multiple sequence alignment of the MUB domains of *Lb. reuteri* shows that they are 90 % identical in sequence, while the first 100 residues of each domain share less than 15 % sequence identity with the second half of the domain (data not shown). This suggests that the copies of the *Lb. reuteri* domain have evolved from one large MUB domain, in turn suggesting that the large domain functions as a biological unit. A possible explanation for this difference in size will be discussed below.

This variability in size makes it difficult to determine the boundaries of MUB domains, a problem often encountered in defining protein domains (Ekman *et al.*, 2005). HMM searches with models based on the MUB domains of different proteins suggest contradicting boundaries (see below). This problem is further complicated by the presence of what seem to be partial MUB domains flanking complete MUB domains in the same protein. Ultimately, we were able to define the boundaries of the MUB domain by comparing different sets of sequences of highly similar proteins that differed in their number of MUB domains: (i) protein L39650 from *Lc. lactis* IL1403 and its orthologue RLCR01214 from *Lc. lactis* SK11, (ii) two highly similar proteins RLBR01191 and RLBR01264 from *Lb. brevis*, and (iii) orthologues of lp_1229 from different *Lb. plantarum* strains (G. Pretzer and others, unpublished data).

Next, we searched the protein databases for MUB domains using domain boundaries derived from multiple sequence alignments of orthologous and paralogous proteins as described above. We identified a total of 48 proteins containing at least one MUB domain in nine different species. Most were lactobacilli that are known inhabitants of the gastrointestinal tract, while others were species commonly used in food fermentations (Altermann *et al.*, 2005; Aryanta *et al.*, 1991; Bolotin *et al.*, 2001; Kleerebezem *et al.*, 2003; Pridmore *et al.*, 2004; Zoetendal *et al.*, 2002). A schematic overview of the 30 proteins with three or more MUB domains is given in Fig. 1. Table 1 lists the species in which proteins containing one or more MUB domains were identified, while a complete list of putative mucus-binding proteins and their predicted features is given in supplementary Table B (available with the online version of this paper). A multiple sequence alignment of selected MUB domains and a predicted secondary structure of the domain can be found in the supplementary Figs A and B (available with the online version of this paper). An HMM based on

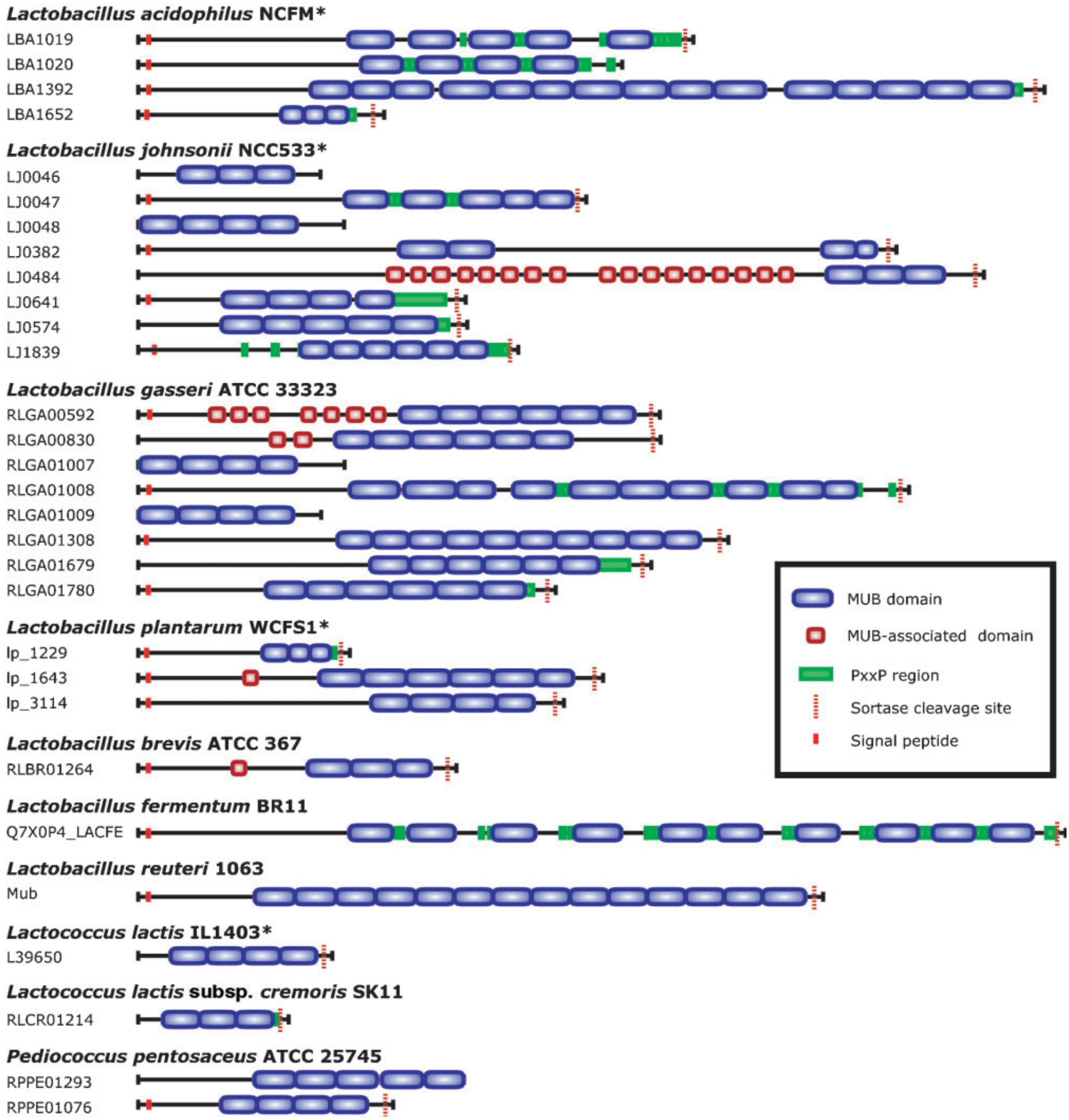
Lactobacillus acidophilus NCFM*
LBA1019
LBA1020
LBA1392
LBA1652
Lactobacillus johnsonii NCC533*
LJ0046
LJ0047
LJ0048
LJ0382
LJ0484
LJ0641
LJ0574
LJ1839
Lactobacillus gasseri ATCC 33323
RLGA00592
RLGA00830
RLGA01007
RLGA01008
RLGA01009
RLGA01308
RLGA01679
RLGA01780
Lactobacillus plantarum WCFS1*
lp_1229
lp_1643
lp_3114
Lactobacillus brevis ATCC 367
RLBR01264
Lactobacillus fermentum BR11
Q7X0P4_LACFE
Lactobacillus reuteri 1063
Mub
Lactococcus lactis IL1403*
L39650
Lactococcus lactis subsp. *cremoris* SK11
RLCR01214
Pediococcus pentosaceus ATCC 25745
RPPE01293
RPPE01076
MUB domain
MUB-associated domain
PxxP region
Sortase cleavage site
Signal peptide

Fig. 1. Domain architecture of proteins with three or more MUB domains. An asterisk indicates a species for which the complete genome sequence is available. The different domains and other sequence features are explained in detail in the text.

Table 1. Species containing at least one protein with one or more MUB domain

Organism	No. of MUB-containing proteins
Lactobacillus gasseri ATCC 33323	13
Lactobacillus acidophilus NCFM	12
Lactobacillus johnsonii NCC533	9
Lactobacillus plantarum WCFS1	4
Lactobacillus reuteri 1063*	2
Lactobacillus brevis ATCC 367	2
Lactobacillus fermentum BR11*	2
Pediococcus pentosaceus ATCC 25745	2
Lactococcus lactis IL1403	1
Lactococcus lactis SK11	1

*Sequences from UniProt (no genome sequence available).

the complete set of MUB domains identified is given in supplementary File 1 (available with the online version of this paper).

MUB-domain-containing proteins are most abundant in lactobacilli that are found mainly in the gastrointestinal tract, supporting the hypothesis that the domain is involved in adherence to intestinal mucus. The genomes of bacteria that have a broader lifestyle and are less frequently encountered in the gastrointestinal tract, such as *Lb. plantarum* (Kleerebezem *et al.*, 2003), encode a smaller number of these proteins. Compared to lactobacilli of the gastrointestinal tract, 'domesticated' *Lc. lactis* strains live in a more restricted habitat (Bolotin *et al.*, 2001), which could explain the presence of only a single MUB-domain-containing protein in this bacterium.

Comparison of MUB and Pfam-MucBP domains

The model MucBP from the Pfam database is similar to part of the MUB domain described in this study. The Pfam model describes a sequence of approximately 50 aa, while we predict the MUB domain to be approximately 200 residues in length.

The MucBP model tends to either leave large gaps between the different copies of the MUB domain in many proteins or predict the presence of multiple instances of the domain in situations where only a single domain is present. Also, the phylogenetic distribution of the Pfam-MucBP domain is much broader than that of the domain described in this study; it is found in many proteins of *Listeria monocytogenes* and even in a single protein supposedly from *Homo sapiens*, although the presence in this protein of both a Gram-positive signal peptide and an LPxTG-type sortase anchor suggest that this protein is in fact of bacterial origin. In our opinion, the relatively weak similarity between these domains and the *Lb. plantarum* MUB domain that has been experimentally shown to have mannose-binding properties or the MUB domains of the protein Mub from *Lb. reuteri* does not warrant the inclusion of these *Listeria* proteins in the set of putative MUB proteins discussed here. Many proteins with significant hits to the Pfam-MucBP domain, but with no significant hits to the MUB domain of LAB, contain multiple copies of the leucine-rich repeat domain (a domain thought to be involved in protein–protein interactions; Kobe & Kajava, 2001). This observation does not hold true for the putative MUB proteins we have identified, indicating that they might have different roles.

The difference in size between MUB and MucBP can in part be explained by the distinct N-terminal region of the MUB domain that we have identified; this part is present in 43 of the 48 proteins containing the MUB domain. Fig. 2 shows a multiple alignment of the N terminus of a subset of domains. This part of the domain does not merely act as a spacer or a flexible region since it has numerous conserved residues and is predicted to contain distinct secondary structure elements. As shown in the alignment, in some cases the N-terminal part of the MUB domain is separated from the rest of the domain by a PxxP region (discussed in more detail below). The N-terminal part of the MUB domain is never found without the C-terminal part, showing that it is in fact part of the MUB domain and not functioning as a separate domain.

Putative mucus-binding proteins and genome context

The physical proximity of genes with linked functions can offer an organism a selective advantage, making such gene clusters less prone to break-up during evolution than others. Therefore, conservation of gene context can be an indicator of linked function and interaction of encoded proteins (Dandekar *et al.*, 1998; Marcotte *et al.*, 1999). Gene clusters encoding MUB proteins were found to be only conserved over relatively short evolutionary distances: conserved clusters were only detected in *Lactobacillus johnsonii*, *Lactobacillus gasseri* and *Lactobacillus acidophilus*, three species which are closely related (Altermann *et al.*, 2005; Pridmore *et al.*, 2004). This lack of conservation over larger phylogenetic distances is in good agreement with the observation that even bacteria of the same genus, such as *Lb. plantarum* and *Lb. johnsonii*, have their own set of extracellular proteins (Boekhorst *et al.*, 2004). Although the exact context of genes encoding MUB-domain proteins is not conserved, these proteins are often encoded in gene clusters together with other putative extracellular proteins (based on the presence of a signal peptide), suggesting that these extracellular proteins have a functional relation. In several cases multiple proteins containing the MUB domain are encoded in a single gene cluster; Fig. 3 shows an example of a cluster of MUB-containing proteins in *Lb. acidophilus* and a different conserved cluster found in both *Lb. gasseri* and *Lb. johnsonii*.

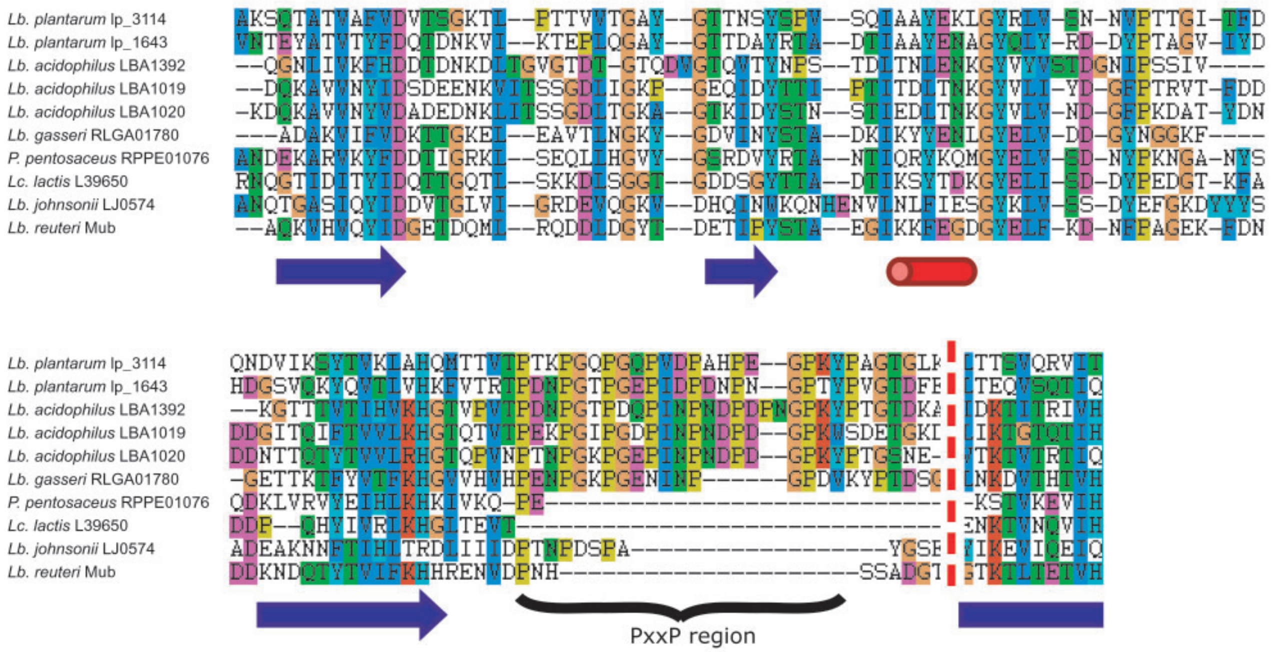

Fig. 2. Multiple alignment of the N-terminal part of selected MUB domains. Blue arrows indicate predicted beta strands; the red cylinder indicates a predicted α-helix. The dashed vertical line indicates the end of the N-terminal part of the domain. The alignment was visualized with CLUSTALX using the default colouring scheme (Thompson *et al*., 1997).

MUB sequence conservation and binding specificity

In most cases, the MUB domains of a single protein are more similar to each other than to the MUB domains in other proteins of the same species. This suggests that the introduction of multiple copies of the domain in a single protein occurred after speciation. In a few proteins, two different versions of the MUB domain can be distinguished; experiments by Roos & Jonsson (2002) show that the different MUB domain types can have different adhesion targets, suggesting a broadening of the range of mucus components such a protein can adhere to. In situations where all the copies of the MUB domains in a protein are highly similar, the role of a larger number of domains could be an increased affinity to mucins.

The high variability in the number of MUB domains in related proteins, even in orthologous proteins from closely related species, exemplifies the relative ease with which the domain is duplicated or deleted in evolution. As an example, Fig. 4 displays two orthologous proteins from *Lb. plantarum* and *Lb. brevis*, which have 6 and 3 homologous MUB domains, respectively. A possible evolutionary scenario that would explain the tree is an ancestral protein containing three copies of the MUB domain followed by a single duplication of the first MUB domain and two successive duplications of the second MUB domain in the *Lb. plantarum* version of the ancestral protein. The relatively frequent deletion and duplication of the MUB domain might be explained by repetitive DNA sequence in the boundaries of the domain. However, analysis of the boundaries of the domain did not yield any repetitive structures such as inverted repeats or tandem repeats.

Cell wall anchors and signal peptides

Of the 30 proteins containing three or more copies of the MUB domain, 19 are predicted to contain a signal peptide (Fig. 1). A signal peptide is an N-terminal signal sequence that targets a protein to the bacterial cell wall (von Heijne, 1989). For 5 of the residual 11 proteins, originally predicted not to contain a signal peptide, we were able to identify a signal peptide by selecting an alternative translation start site or by the introduction of a single frameshift (see supplementary Table B, available with the online version of this paper). Frameshifts could either be a sequencing artefact or a true frameshift; the latter suggests the gene containing the frameshift does not encode a functional protein. In addition to a signal peptide, most of the proteins with multiple MUB domains contain a C-terminal anchoring motif, called LPxTG, that is recognized by a family of enzymes called sortases for covalent attachment to the peptidoglycan of the bacterial cell wall (Navarre & Schneewind, 1999). The presence or absence of signal peptides and LPxTG-like motifs is summarized in supplementary Table A (available with the online version of this paper).

The MUB-domain-containing proteins without an LPxTG anchor or a signal peptide are often encoded next to proteins

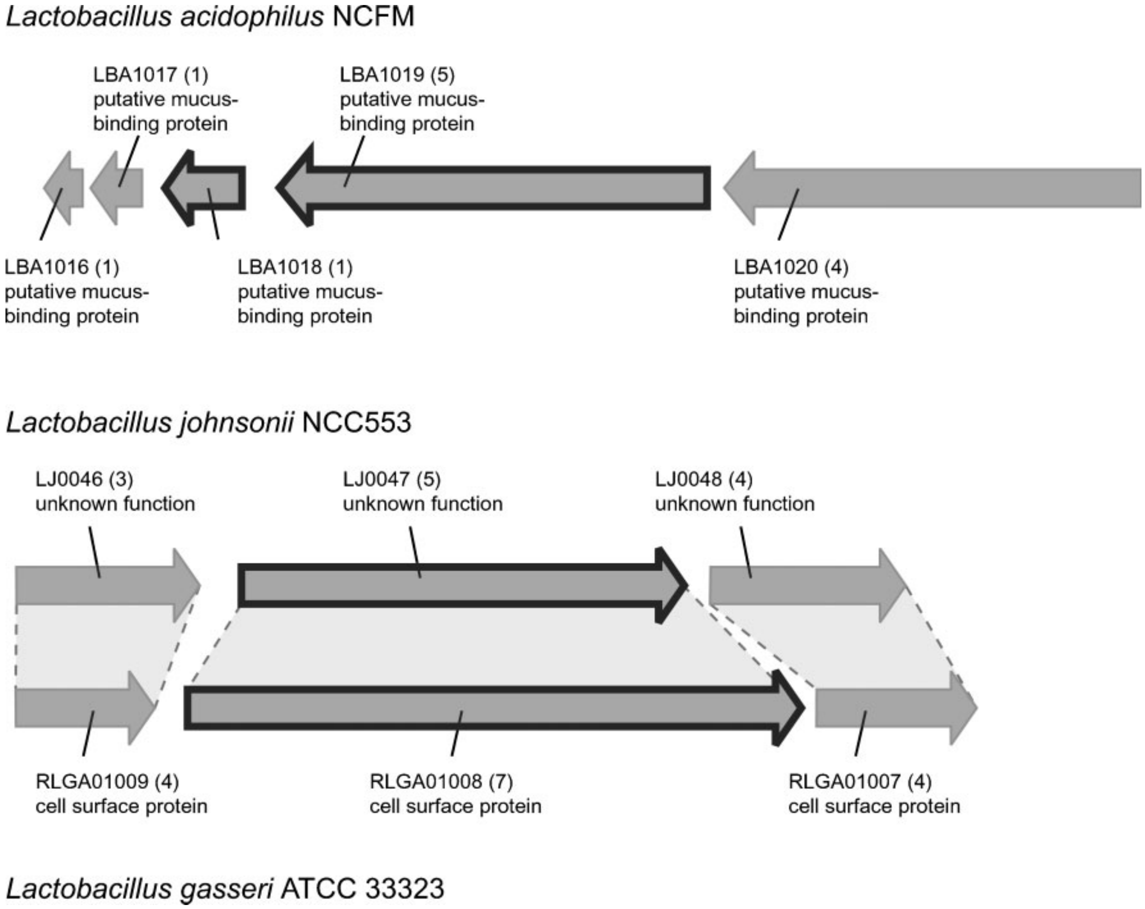
Lactobacillus acidophilus NCFM
LBA1017 (1) putative mucus-binding protein
LBA1019 (5) putative mucus-binding protein
LBA1016 (1) putative mucus-binding protein
LBA1018 (1) putative mucus-binding protein
LBA1020 (4) putative mucus-binding protein
Lactobacillus johnsonii NCC553
LJ0046 (3) unknown function
LJ0047 (5) unknown function
LJ0048 (4) unknown function
RLGA01009 (4) cell surface protein
RLGA01008 (7) cell surface protein
RLGA01007 (4) cell surface protein
Lactobacillus gasseri ATCC 33323

Fig. 3. Gene clusters encoding cell-surface proteins. The numbers in parentheses indicate the number of MUB domains found. Arrows with a thick black outline represent genes encoding proteins that are predicted to contain an LPxTG-like membrane anchor. The grey bars connect orthologous genes.

containing the MUB domain, a signal peptide and an LPxTG-like motif. They could either be non-functional remnants of extracellular proteins or function at the bacterial cell wall through some other mechanism, e.g. interaction with extracellular proteins that do contain a membrane anchor.

PxxP regions

Many MUB proteins contain proline-rich amino acid stretches, designated PxxP regions. In these regions, proline residues are separated by two (in rare cases 1 or 3) non-proline residues. Sequences of at least 13 residues in length

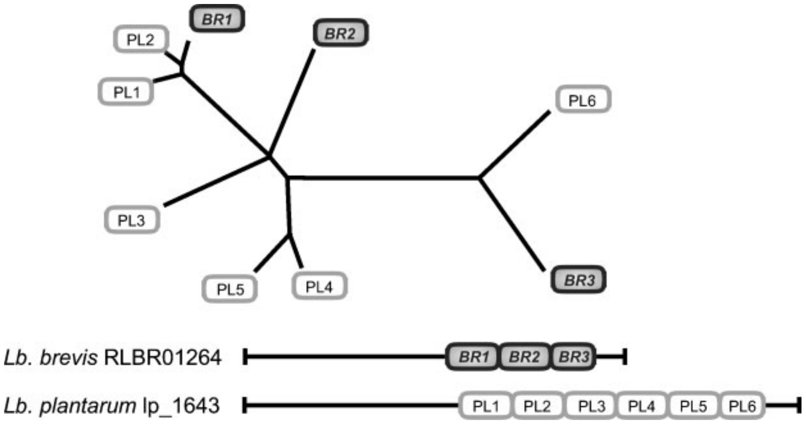

Fig. 4. Unrooted phylogenetic tree of the individual MUB domains of the orthologous proteins lp_1643 (*Lb. plantarum*) and RLBR01264 (*Lb. brevis*). The *Lb. plantarum* protein contains six MUB domains, while the *Lb. brevis* protein contains only three. The N-terminal parts of the two proteins are over 70 % identical, while the identity between MUB domains varies between 33 and 90 %, with a mean of 52 %.

containing at least five P residues were considered PxxP regions. About half of the putative mucus-binding proteins we identified contained at least one PxxP region, always inserted in or flanking an MUB domain. In some of the cases where such a region is found inside an MUB domain, only a subset of the MUB domains of a specific protein contain a PxxP region. Combined with the observation that the MUB domains of such a protein are often more similar to each other than to any other MUB domain, this suggests that the insertion or deletion of PxxP regions are quite common events.

It has been suggested that proline residues allow a polypeptide chain to make sharp bends or twists (Fischetti, 2000). In this scenario, the function of the PxxP regions could be to generate flexibility of the protein chain. The presence of a PxxP region between the C-terminal membrane anchor and the last MUB domain of many of the MUB proteins supports this putative function of the PxxP regions (Fig. 1).

In eukaryotes, proline-rich motifs are known to be involved in binding to so-called SH3 domains (described by Pfam entries SH3_1 and SH3_2). In these interactions, the proline-rich sequence forms a polyproline type II helical conformation that fits into the hydrophobic groove of the SH3 domain (Agrawal & Kishan, 2002). The proline-rich motifs bound by the SH3 domain are approximately 13 residues long; the difference in size compared to the PxxP regions found in the MUB proteins, which in some cases reach lengths of over 50 residues, suggests that the bacterial regions could form similar secondary structure elements, but might have a different function.

Identification of other domains of MUB proteins

To gain further insight into the presence and putative function of other domains of mucus-binding proteins, generally preceding the MUB domains, all proteins containing one or more MUB domain were scanned with HMMs from the Pfam, Superfam and SMART databases. In many cases we identified sequences in the N-terminal region of these proteins which are similar to either binding domains or enzymic domains found in other extracellular proteins, such as glucanase and pectin lyase-like domains (supplementary Table B, available with the online version of this paper). However, in most cases these regions scored just below the threshold suggested for the various HMMs, indicating that these potential domains have similar, but not identical, enzymic functions. The similarity to glucanase and pectin lyase-like domains suggests that these putative domains may be involved in degradation of complex polysaccharides or mucus-associated glycosylation moieties.

In addition to domains with similarity to domains with a known function or structure, we identified a previously undescribed domain of approximately 70 aa in size. This domain, which we call MUB-associated domain (Mubad), is present in six of the proteins containing the MUB domain and the number of Mubad domains per protein varies between 1 and 18 (Fig. 1). A multiple sequence alignment of Mubad domains (supplementary Fig. C, available with the online version of this paper) shows that this domain is not highly conserved. However, the presence of Mubad domains only in proteins that contain MUB domains suggests that the association is significant. Comparison of an HMM based on a multiple alignment of the Mubad domain to models from the Pfam and Superfam databases did not detect any known domains homologous to the Mubad domain. Although the function of this Mubad domain remains unclear, it again illustrates the complex domain architecture of the putative mucus-binding proteins.

Concluding remarks

The MUB domain is very variable in size and sequence, making it difficult to determine precise domain boundaries. The use of orthologous proteins with different numbers of MUB domains allowed us to identify putative functional units. The high variability in the number of MUB domains in putative mucus-binding proteins suggests that the MUB domain is often duplicated or deleted in evolution. In contrast to the MucBP domain from the Pfam database, the MUB domain appears to be only present in LAB, with the highest abundance in lactobacilli of the gastrointestinal tract, strongly suggesting that the MUB domain is a functional unit specific to LAB that could fulfil an important function in host–microbe interactions.

The genomes sequenced from intestinal bacteria are presently biased towards LAB, due to the relevance of LAB to food and health. In the future, the ever-increasing number of available genome sequences might lead to the identification of MUB-domain-containing proteins in other species and other types of mucus-binding domains.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Netherlands Organization for Scientific Research (NWO-BMI project 050.50.206).

REFERENCES

Agrawal, V. & Kishan, K. V. (2002). Promiscuous binding nature of SH3 domains to their target proteins. *Protein Pept Lett* **9**, 185–193.

Altermann, E., Russell, W. M., Azcarate-Peril, M. A. & 11 other authors (2005). Complete genome sequence of the probiotic lactic acid bacterium *Lactobacillus acidophilus* NCFM. *Proc Natl Acad Sci U S A* **102**, 3906–3912.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

Aryanta, R. W., Fleet, G. H. & Buckle, K. A. (1991). The occurrence and growth of microorganisms during the fermentation of fish sausage. *Int J Food Microbiol* **13**, 143–155.

Bailey, T. L. & 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.

Bairoch, A., Apweiler, R., Wu, C. H. & 12 other authors (2005). The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33**, D154–D159.

Bateman, A., Coin, L., Durbin, R. & 10 other authors (2004). The Pfam protein families database. *Nucleic Acids Res* **32**, D138–D141.

Bayliss, C. E. & Houston, A. P. (1984). Characterization of plant polysaccharide- and mucin-fermenting anaerobic bacteria from human feces. *Appl Environ Microbiol* **48**, 626–632.

Boekhorst, J., Siezen, R. J., Zwahlen, M. C. & 7 other authors (2004). The complete genomes of *Lactobacillus plantarum* and *Lactobacillus johnsonii* reveal extensive differences in chromosome organization and gene content. *Microbiology* **150**, 3601–3611.

Boekhorst, J., de Been, M. W., Kleerebezem, M. & Siezen, R. J. (2005). Genome-wide detection and analysis of cell wall-bound proteins with LPxTG-like sorting motifs. *J Bacteriol* **187**, 4928–4934.

Bolotin, A., Wincker, P., Mauger, S., Jaillon, O., Malarme, K., Weissenbach, J., Ehrlich, S. D. & Sorokin, A. (2001). The complete genome sequence of the lactic acid bacterium *Lactococcus lactis* ssp. lactis IL1403. *Genome Res* **11**, 731–753.

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett* **286**, 47–54.

Coconnier, M. H., Klaenhammer, T. R., Kerneis, S., Bernet, M. F. & Servin, A. L. (1992). Protein-mediated adhesion of *Lactobacillus acidophilus* BG2FO4 on human enterocyte and mucus-secreting cell lines in culture. *Appl Environ Microbiol* **58**, 2034–2039.

Conway, P. L. & Kjelleberg, S. (1989). Protein-mediated adhesion of *Lactobacillus fermentum* strain 737 to mouse stomach squamous epithelium. *J Gen Microbiol* **135**, 1175–1186.

Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **23**, 324–328.

Dekker, J., Rossen, J. W., Buller, H. A. & Einerhand, A. W. (2002). The MUC family: an obituary. *Trends Biochem Sci* **27**, 126–131.

Doolittle, R. F. & Bork, P. (1993). Evolutionarily mobile modules in proteins. *Sci Am* **269**, 50–56.

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

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E. & Relman, D. A. (2005). Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113.

Ekman, D., Bjorklund, A. K., Frey-Skott, J. & Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol* **348**, 231–243.

Fischetti, V. A. (2000). In *Gram-Positive Pathogens*, pp. 11–24. Washington, DC: American Society for Microbiology.

Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**, 903–919.

Hooper, L. V. & Gordon, J. I. (2001). Commensal host–bacterial relationships in the gut. *Science* **292**, 1115–1118.

Kleerebezem, M., Boekhorst, J., van Kranenburg, R. & 17 other authors (2003). Complete genome sequence of *Lactobacillus plantarum* WCFS1. *Proc Natl Acad Sci U S A* **100**, 1990–1995.

Kobe, B. & Kajava, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* **11**, 725–732.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P. & Bork, P. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32**, D142–D144.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753.

McGuffin, L. J., Bryson, K. & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405.

Navarre, W. W. & Schneewind, O. (1999). Surface proteins of Gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol Mol Biol Rev* **63**, 174–229.

Overbeek, R., Larsen, N., Walunas, T. & 19 other authors (2003). The ERGO genome analysis and discovery system. *Nucleic Acids Res* **31**, 164–171.

Pretzer, G., Snel, J., Molenaar, D. & 7 other authors (2005). Biodiversity-based identification and functional characterization of the mannose-specific adhesin of *Lactobacillus plantarum. J Bacteriol* **187**, 6128–6136.

Pridmore, D., Berger, B., Desiere, F. & 12 other authors (2004). The genome sequence of the probiotic intestinal bacterium *Lactobacillus johnsonii* NCC 533. *Proc Natl Acad Sci U S A* **101**, 2512–2517.

Reid, G., Sanders, M. E., Gaskins, H. R. & 7 other authors (2003). New scientific paradigms for probiotics and prebiotics. *J Clin Gastroenterol* **37**, 105–118.

Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277.

Roos, S. & Jonsson, H. (2002). A high-molecular-mass cell-surface protein from *Lactobacillus reuteri* 1063 adheres to mucus components. *Microbiology* **148**, 433–442.

SantaLucia, J., Jr (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* **95**, 1460–1465.

Servin, A. L. (2004). Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. *FEMS Microbiol Rev* **28**, 405–440.

Soding, J. (2005). Protein homology detection by HMM–HMM comparison. *Bioinformatics* **21**, 951–960.

Sonnenburg, J. L., Xu, J., Leip, D. D., Chen, C. H., Westover, B. P., Weatherford, J., Buhler, J. D. & Gordon, J. I. (2005). Glycan foraging *in vivo* by an intestine-adapted bacterial symbiont. *Science* **307**, 1955–1959.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & 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.

Ton-That, H., Marraffini, L. A. & Schneewind, O. (2004). Protein sorting to the cell wall envelope of Gram-positive bacteria. *Biochim Biophys Acta* **1694**, 269–278.

von Heijne, G. (1989). The structure of signal peptides from bacterial lipoproteins. *Protein Eng* **2**, 531–534.

Zoetendal, E. G., von Wright, A., Vilpponen-Salmela, T., Ben-Amor, K., Akkermans, A. D. & de Vos, W. M. (2002). Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces. *Appl Environ Microbiol* **68**, 3401–3407.