Microbiology (2006), **152**, 3175–3183 **DOI** 10.1099/mic.0.29217-0

The predicted secretome of *Lactobacillus plantarum* WCFS1 sheds light on interactions with its environment

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

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

[1]Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, 6500HB Nijmegen, The Netherlands

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

[3]NIZO food research, Ede, The Netherlands

Received 14 June 2006
Revised 15 August 2006
Accepted 17 August 2006

The predicted extracellular proteins of the bacterium *Lactobacillus plantarum* were analysed to gain insight into the mechanisms underlying interactions of this bacterium with its environment. Extracellular proteins play important roles in processes ranging from probiotic effects in the gastrointestinal tract to degradation of complex extracellular carbon sources such as those found in plant materials, and they have a primary role in the adaptation of a bacterium to changing environmental conditions. The functional annotation of extracellular proteins was improved using a wide variety of bioinformatics methods, including domain analysis and phylogenetic profiling. At least 12 proteins are predicted to be directly involved in adherence to host components such as collagen and mucin, and about 30 extracellular enzymes, mainly hydrolases and transglycosylases, might play a role in the degradation of substrates by *L. plantarum* to sustain its growth in different environmental niches. A comprehensive overview of all predicted extracellular proteins, their domains composition and their predicted function is provided through a database at http://www.cmbi.ru.nl/secretome, which could serve as a basis for targeted experimental studies into the function of extracellular proteins.

INTRODUCTION

Lactobacillus plantarum is a versatile and widespread micro-organism found in environments ranging from vegetable, dairy and meat fermentations to the human gastrointestinal (GI) tract (Kleerebezem *et al.*, 2003). Some strains are marketed as probiotics that are claimed to provide a health benefit for the consumer through interactions with the human GI system (for a review see de Vries *et al.*, 2005). Attachment of probiotic bacteria to specific sites on intestinal mucosa cells might lead to competitive exclusion of pathogens and/or modulation of host cell responses. Proteins that are exposed on the bacterial cell surface are considered to play an important role in such interactions. After secretion, many extracellular proteins are attached covalently or non-covalently to components of the bacterial cell wall, such as peptidoglycan or teichoic acids. Others are anchored to the membrane through one or more membrane-spanning helices or covalent coupling to lipids. At the cell surface these proteins are involved in processes such as signal transduction, recognition, binding and degradation of complex nutrients (e.g. polysaccharides), nutrient uptake and adherence to host cells (Buck *et al.*, 2005; Kawai *et al.*, 2006; Roos & Jonsson, 2002). For example, the protein Msa of *L. plantarum* has been shown to be a mannose-specific adhesin (Pretzer *et al.*, 2005).

Extracellular proteins are often large proteins consisting of many modules or domains (Bork, 1991). The modification and recombination of functional modules plays an important role in the evolution of protein function (Doolittle & Bork, 1993). A number of extracellular proteins contain so-called domain repeats – multiple copies of similar domains separated by 30 or fewer amino acids – which are often involved in binding (Boekhorst *et al.*, 2006; Cabanes *et al.*, 2002; Zhang *et al.*, 2000). In addition to domain repeats, some extracellular proteins contain regions of low complexity with a repetitive nature; examples are the SD repeat of the adhesion factor ClfB of *Staphylococcus aureus* (Ni Eidhin *et al.*, 1998) and the PxxP regions of putative mucus-binding proteins (Boekhorst *et al.*, 2006). Predicting the function of these proteins is often quite a challenge, because of their complex domain architecture and the relatively poor sequence conservation of extracellular

Abbreviations: GI, gastrointestinal; HMM, hidden Markov model.

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

proteins among different species of bacteria. The identification and characterization of domains and repeats can play an important role in elucidating the function of extracellular proteins.

Here we describe and analyse the predicted extracellular proteins (or 'secretome') of *L. plantarum* to provide insight into possible mechanisms of interactions of this species with its variable environment. We present an improved prediction of function of these proteins by a combination of different approaches: (i) by analysing the occurrence of known domains, (ii) by identification of new domains and repeats and (iii) by studying the phylogenetic distribution of homologues of the secreted proteins. The relevance of the predicted functions of the secretome proteins in relation to the lifestyle of *L. plantarum* is discussed. A database providing updated annotation of the secreted proteins of *L. plantarum* and the domain composition of these proteins can be found at http://www.cmbi.ru.nl/secretome.

METHODS

Sequence analysis. Sequence information was obtained from the NCBI bacterial genome database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Sequence similarity was detected by the Smith and Waterman method (Smith & Waterman, 1981) on a DeCypher hardware-accelerated system (Active Motif, Carlsbad, CA). Multiple-sequence alignments were made using Muscle (Edgar, 2004) and visualized with CLUSTALX (Thompson *et al.*, 1997). Phylogenetic trees were reconstructed with PhyML (Guindon & Gascuel, 2003). Creation of hidden Markov models (HMMs) and scanning protein databases with HMMs was done with the HMMER package (Durbin *et al.*, 1998). HMMs were compared using HHSEARCH (Soding, 2005). Known domains were obtained from the Pfam (Bateman *et al.*, 2004) and SUPERFAM (Gough *et al.*, 2001) databases. Low-complexity regions and repeats were identified with SAPS (Brendel *et al.*, 1992). Operons were predicted as described by Wels *et al.* (2006).

Identification of extracellular proteins. Extracellular proteins were predicted on the basis of the presence of a Sec-pathway-dependent (lipoprotein) signal peptide. Class I signal peptides were detected using SignalP3 (Bendtsen *et al.*, 2004). Lipoprotein signal peptides were detected by the consensus sequence L(A,S)(A,G)C at the end of the N-terminal hydrophobic region (Sutcliffe & Russell, 1995). Proteins containing three or more transmembrane helices as determined with TMHMM2 (Krogh *et al.*, 2001) were considered integral membrane proteins and excluded from further analysis. Information on potential sortase recognition sites for anchoring to peptidoglycan was taken from Boekhorst *et al.* (2005). Potential N-terminal transmembrane anchors were predicted by identifying proteins with a predicted signal peptide, but without a good cleavage site for the signal peptidase as determined using SignalP3 (Bendtsen *et al.*, 2004). Potential C-terminal transmembrane anchors were identified based on the presence of a C-terminal transmembrane helix as predicted by TMHMM, flanked by positively charged residues and the absence of a known cleavage site (e.g. sortase recognition sites). Proteins secreted by other mechanisms (e.g. GG-type leader peptides) were not considered.

Identification of novel protein domains. Novel conserved domains were identified by scanning all parts of the putative secreted proteins of *L. plantarum* that do not overlap with known domains for sequence homology. Sequences from homologous regions were aligned and used to build HMMs to search for further occurrences and iterative improvement of HMMs.

Repeated domains within single proteins were identified using the MEME motif discovery tool (Bailey & Elkan, 1994). Conserved regions were aligned automatically followed by manual refinement.

Conserved gene clusters and phylogenetic distributions. Conserved clusters of genes predicted to encode extracellular proteins were analysed using String (von Mering *et al.*, 2005), as was the phylogenetic distribution of homologues of *L. plantarum* secreted proteins. Only predicted functional associations with a score of at least 0·5 (defined by String to be medium–high confidence or better) were considered.

In addition, secreted proteins were grouped into sets with similar phylogenetic distribution patterns on the basis of the presence or absence of homologous proteins in 350 complete genome sequences. The complex domain architecture of many extracellular proteins makes it difficult to identify homologues of *L. plantarum* proteins in other species. Therefore, stringent criteria were used, i.e. only proteins with Smith and Waterman *E*-values below 1×10^{-25} and with at least 60 % sequence overlap were considered to be homologues, in order to prevent proteins sharing only part of their domain composition from being detected as homologues. Sets with similar distribution patterns were chosen based on species subsets (e.g. Gram-positives, *Lactobacillus*-specific).

RESULTS AND DISCUSSION

Extracellular proteins and anchor types

The *L. plantarum* genome is predicted to encode 223 extracellular proteins, of which the large majority have a motif or domain for attachment to the cell surface (Table 1). Fig. 1 illustrates the different types of extracellular proteins and their anchoring mechanisms. Forty-eight of these proteins contain an N-terminal lipobox, a common mechanism for secretion and membrane attachment of proteins through covalent binding of a conserved cysteine residue to a lipid (Sutcliffe & Harrington, 2002). In *L. plantarum* this anchor type is found mainly in substrate-binding proteins of ABC transporters. Twenty-seven proteins contain a C-terminal LPxTG recognition signal (Boekhorst *et al.*, 2005) for covalent attachment to peptidoglycan by sortase, while 10 proteins are predicted to contain one or more copies of the LysM domain, which is thought to be involved in interaction with peptidoglycan (Bateman & Bycroft, 2000; Steen *et al.*, 2003). Of the residual extracellular proteins, 71 contain a predicted N-terminal signal peptide that appears to lack a clearly identifiable signal peptidase cleavage site, suggesting their N-terminal anchoring in the membrane. Ten additional proteins are predicted to be anchored through a C-terminal transmembrane anchor. Finally, the remaining 57 proteins are predicted to be either unattached (i.e. secreted) or associated with the cell wall using other (unknown) mechanisms.

Extracellular proteins were divided into groups reflecting predicted protein function, based on both functional annotation and domain composition. Proteins were

Table 1. Functional classification and anchor types

Functional class	LPxTG anchor	LysM domain	Lipobox	N-terminal membrane anchor	C-terminal membrane anchor	Unknown/secreted*	Total
Adherence	10	0	0	0	0	2	12
Enzyme	1	8	10	34	0	16	69
Phage	0	2	3	0	0	2	7
Regulator	0	0	0	5	0	0	5
Transporter	0	0	22	5	0	3	30
Unknown	16	0	13	27	10	34	100
Total	27	10	48	71	10	57	223

*Includes both secreted proteins and proteins that are associated with the bacterial cell surface through other mechanisms.

classified as enzymes, transporters, regulators, phage-related proteins, adherence proteins (predicted to bind to extracellular macromolecules such as mucins and fibronectin) or unknown (Table 1). Additional listing and classification information of all *L. plantarum* genes predicted to belong to the secretome are provided as supplementary material with the online version of this paper (the supplementary tables can also be found under 'summary' at http://www.cmbi.ru.nl/secretome): Table S1 lists all proteins predicted to be extracellular, Tables S2–S13 list proteins classified by function and anchor type and Tables S14–S17 list identified Pfam domains. An *L. plantarum* secretome database with all predicted extracellular proteins, their classification, predicted function and domain composition can be found at http://www.cmbi.ru.nl/secretome.

An example of the multi-domain compositions of extracellular proteins is provided by the group of proteins containing a LysM domain for anchoring to peptidoglycan (Fig. 2). In addition to the LysM domain, all these proteins contain a domain predicted to have an enzymic function related to the biosynthesis or degradation of polysaccharides, indicating that different enzymes have a common method for anchoring to peptidoglycan.

Enzymes. Extracellular enzymes play a role in secretion and modification of proteins, in maintenance of the bacterial cell wall, and in the modification and degradation of extracellular compounds, allowing for the use of such molecules as a source of nutrients.

We identified 69 extracellular proteins of *L. plantarum* predicted to have an enzymic function (excluding phage proteins), and these were subdivided into functional groups (Table S3). The first group of 40 proteins contains enzymes with a known biological role in processes such as cell-wall

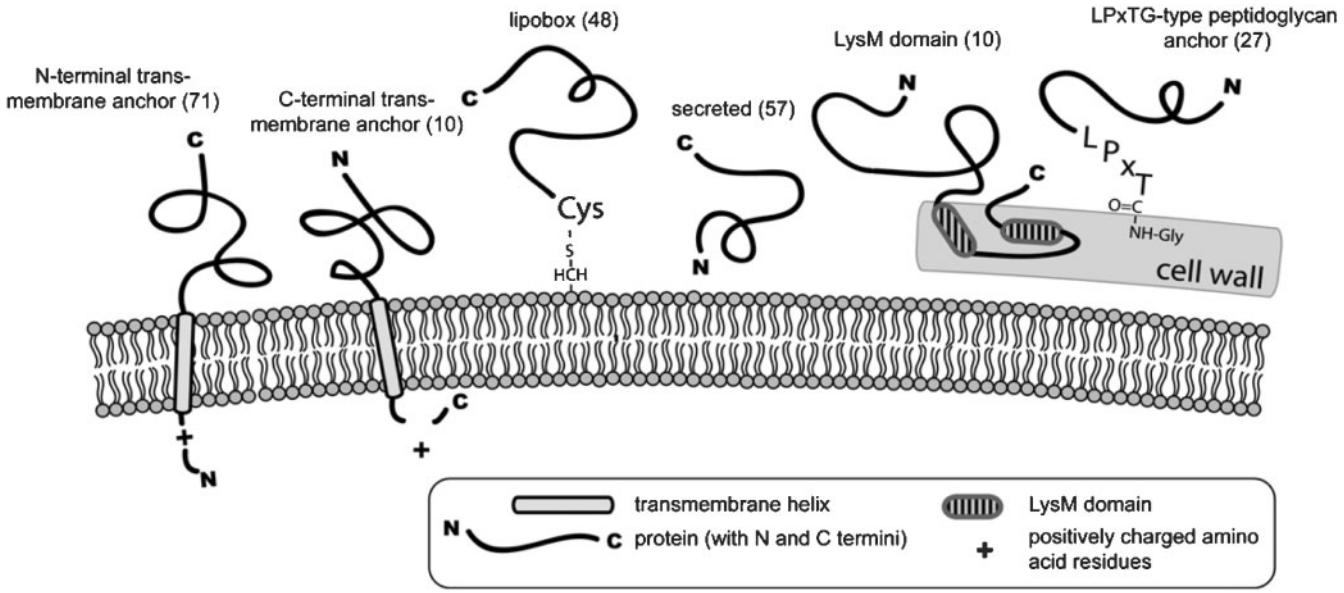

Fig. 1. Extracellular anchored and secreted proteins of *L. plantarum*. Numbers in parentheses are the number of predicted proteins of the different types.

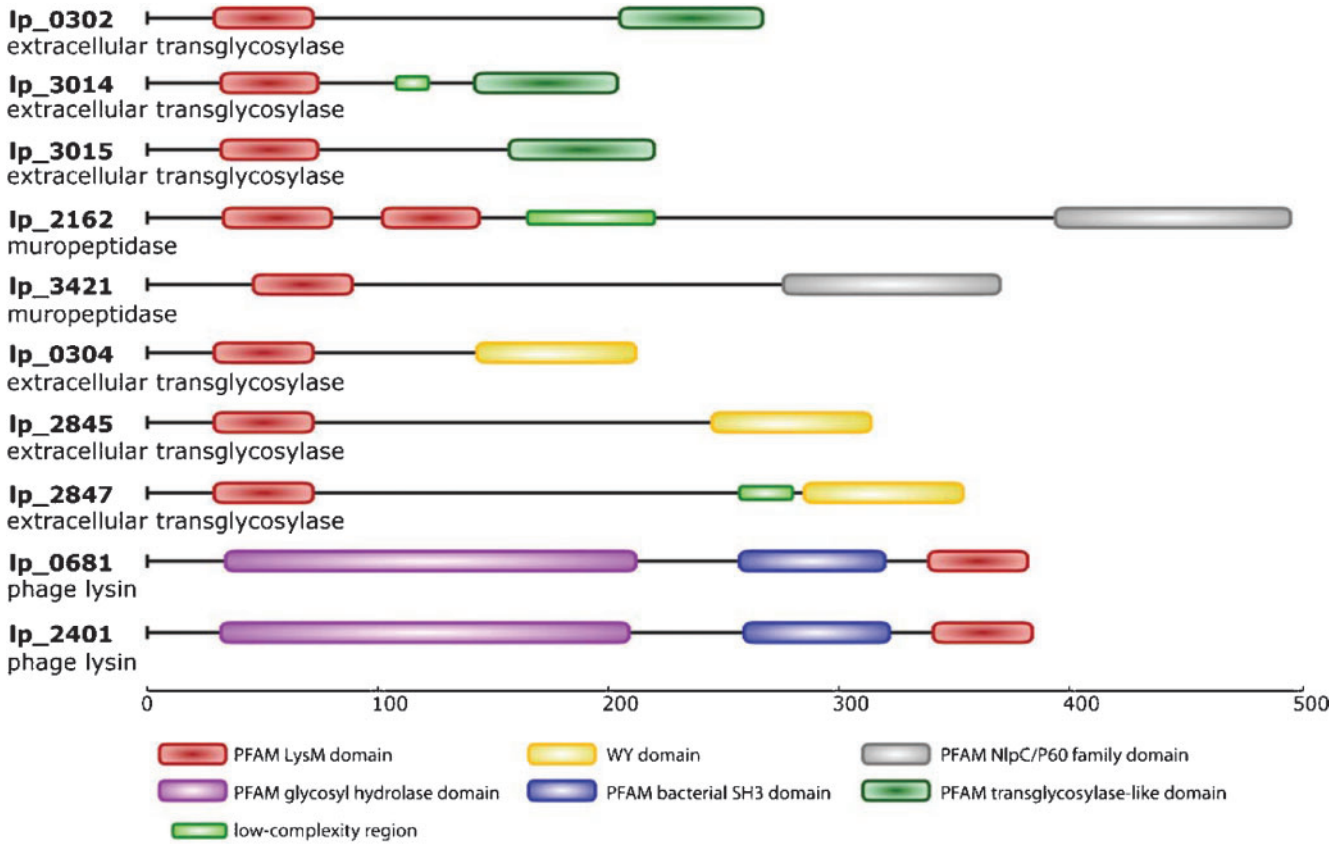

Fig. 2. Domain composition of *L. plantarum* proteins predicted to be associated with the cell wall through LysM domains. Putative protein functions are listed below the ORF names.

biosynthesis and turnover (30 members) and the protein secretion and modification machinery (5 members). These are highly conserved bacterial enzymes required for growth and protein secretion.

The second group of enzymes represents less well-described functions and is expected to include proteins playing a role in the adaptation of *L. plantarum* to specific environments such as the GI tract. In most cases we were able to predict only the general type of reaction catalysed, based on domain architecture or putative conserved catalytic residues, but the specific function they fulfil remains unclear. This group includes predicted metalloproteinases (4), hydrolases such as lipases or esterases (16), and transglycosylases (7). A polysaccharide deacetylase, encoded by a gene flanking a region predicted to encode prophage-related proteins, might be involved in plant-cell-wall degradation.

Adherence proteins and binding domains. Adhesion factors are generally considered to play an important role in host–microbe interactions, and adhesion factors identified in pathogenic bacteria have been shown to play a key role in virulence (Hammerschmidt, 2006; Navarre & Schneewind, 1999). Analogously, in probiotic bacteria they are expected to play a role in persistence and competitive exclusion of pathogens or other health-stimulatory interactions (Marco *et al.*, 2006). The domain composition of the *L. plantarum* proteins predicted to be involved in the adherence to extracellular macromolecules was analysed (Fig. 3). This group consists of three proteins containing domains predicted to be involved in adherence to collagen (Pfam: PF05737) (Symersky *et al.*, 1997), one protein with a chitin-binding domain (Pfam: PF03067) (Yuen *et al.*, 1990), one protein with a fibronectin-binding domain (Pfam: PF05833) (Christie *et al.*, 2002) and seven proteins with domains predicted to be involved in adherence to mucus (Fig. 3). These putative mucus-binding proteins contain either copies of the Pfam MucBP domain and/or copies of the larger MUB domain. While the MucBP domain is not only found in lactic acid bacteria, but also in *Listeria* species, the MUB domain appears to be unique for lactic acid bacteria (Boekhorst *et al.*, 2006). The collagen-binding domain is found in a wide range of firmicutes, while the fibronectin-binding and chitin-binding domains are present in proteins from both eukaryotes and prokaryotes. Of the 12 proteins identified as putative adhesion factors, ten proteins contain an LPxTG-like peptidoglycan anchor. One of these (lp_1229; *msa*) has been experimentally determined to encode a mannose-specific adhesin (Pretzer *et al.*, 2005), which is proposed to fulfil a

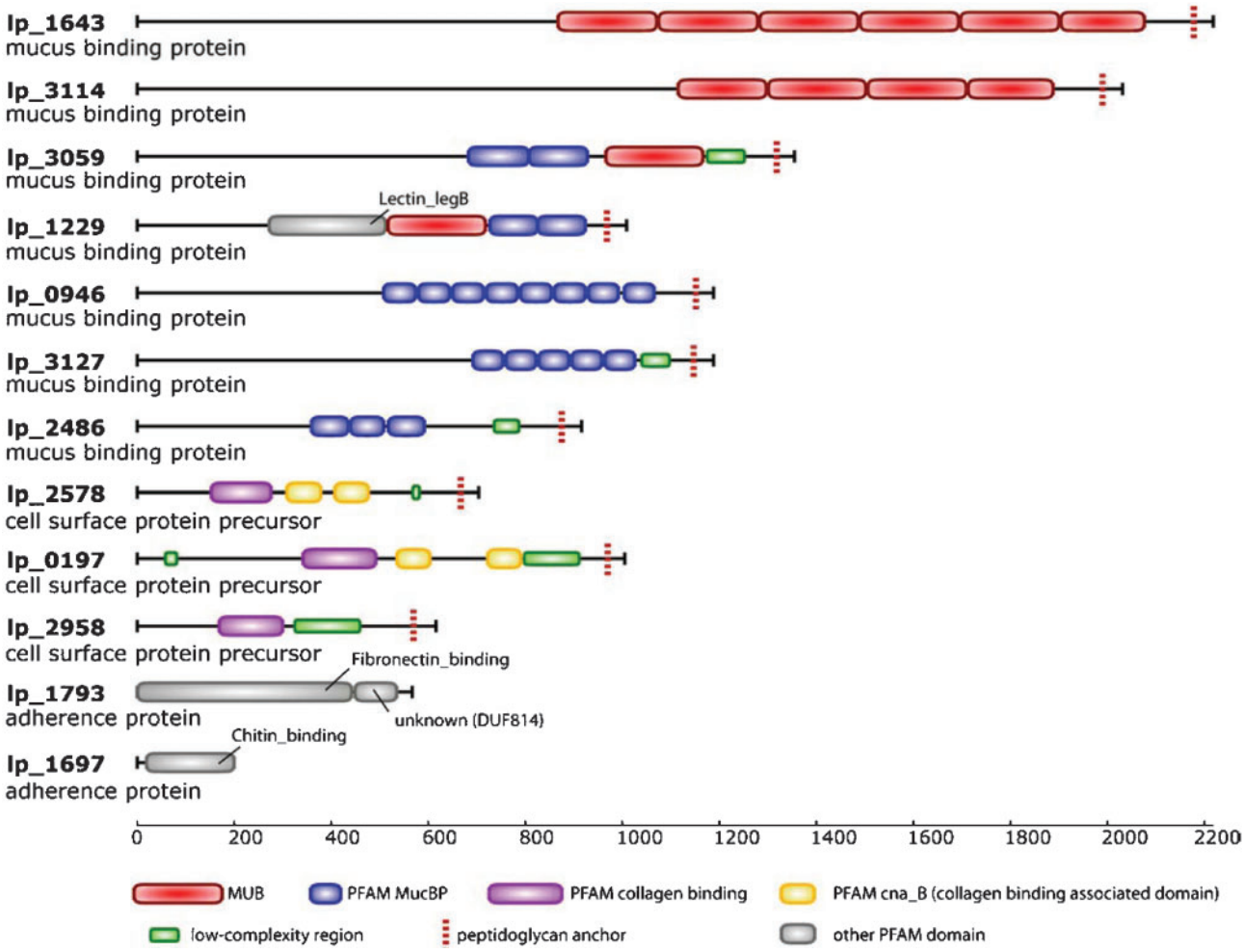

Fig. 3. Domain composition of *L. plantarum* proteins predicted to be involved in the adherence to extracellular macromolecules. Putative protein functions are listed below the ORF names.

key role in reducing pathogenic efficacy of *Escherichia coli* strains through a competitive exclusion mechanism (Adlerberth *et al.*, 1996; Pretzer *et al.*, 2005).

In addition to these putative adhesins, other proteins of *L. plantarum* contain binding domains that are typical for bacterial extracellular proteins (as described in Pfam), such as bacterial immunoglobulin-like domains (Ig-like, PKD), phospholipid-binding domain (F5/8 type C), and also several conserved domains of unknown function (Table S15, Table S17).

Novel domains

In addition to scanning secreted proteins with models from the Pfam database describing known domains, we identified other putative functional regions by scanning for conserved sequence segments that do not overlap with one or more of the known domains. This type of analysis can provide leads for the identification of novel conserved domains, which could enable targeted experimental studies to evaluate their possible function(s). Domains of interest could include those that contain putative conserved catalytic-site-like residues or specific binding pockets that might be involved in the recognition, degradation and/or modification of extracellular substrates.

We identified ten putative domains not yet described in the Pfam database; of these, the four domains called WxL1, WxL2, CSH and WY are present only once per protein (discussed in more detail later), while the other six are repeated domains. Table 2 provides an overview of these putative domains.

One of the repeated domains, the so-called MUB domain, has been discussed in one of our previous studies (Boekhorst *et al.*, 2006) and is present in four *L. plantarum* proteins (Fig. 3). The other five repeated domains, labelled repeat_1 to repeat_5 and ranging in size from 20 to 60 amino acids, represent novel repeated domains. Four of these domains (repeat_1 to repeat_4) are present in single proteins encoded by the *L. plantarum* genome, while the fifth

Table 2. Novel domains

Domain name	Predicted function*	ORFs predicted to contain the domain
MUB	Mucus binding	lp_1229, lp_3114, lp_3059, lp_1643
Repeat_1	Domain of unknown function	lp_0800
Repeat_2	Domain of unknown function	lp_2145
Repeat_3	Domain of unknown function	lp_2925
Repeat_4	Domain of unknown function	lp_3001
Repeat_5	Domain of unknown function	lp_2796, lp_2795, lp_3117, lp_3075, lp_0800
WxL1	Proteins containing this domain are encoded in conserved gene clusters predicted to be involved in carbon source acquisition	lp_3679, lp_3452, lp_3453, lp_3412, lp_3414, lp_1446, lp_3073, lp_1450, lp_2175, lp_1449, lp_2978, lp_3116, lp_3067
WxL2	Proteins containing this domain are encoded in conserved gene clusters predicted to be involved in carbon source acquisition	lp_3450, lp_3064, lp_3676, lp_3075, lp_0297, lp_2173, lp_3117, lp_2975
CSH	Hydrolase	lp_2620, lp_1124, lp_3393, lp_0461, lp_3265
WY	Transglycosylase	lp_0304, lp_3050, lp_2845, lp_2847

*The domains and their predicted functions are described in detail in the main text.

domain (repeat_5), approximately 50 amino acids in length, is found in two to eight copies in five *L. plantarum* proteins. A multiple sequence alignment of repeat_5 (see supplementary Fig. S1, available with the online version of this paper) was used to build an HMM model for this protein element, which was used in comparison to models of structurally related proteins from the SUPERFAMILY database. These analyses revealed that repeat_5 may have an L-domain-like fold. This fold class contains domains involved in protein–protein interactions such as the internalin leucine-rich repeat regions known from *Listeria monocytogenes*. Internalins are known to play a role in host–microbe interactions (Gaillard *et al.*, 1991), suggesting that the *L. plantarum* proteins in which repeat_5 was identified might have a similar function.

The WY domain. Four *L. plantarum* proteins contain a C-terminal domain of approximately 60 amino acids characterized by conserved tryptophan and tyrosine residues. Experimental studies in *Lactobacillus gasseri* and *Lactobacillus johnsonii* suggest that proteins containing this domain are involved in aggregation and/or maintenance of cell shape (Jankovic *et al.*, 2003; Ventura *et al.*, 2002). Comparison of an HMM based on the multiple-sequence alignments of the four WY domains of *L. plantarum* with domains from the Pfam database revealed that the domain displays sequence similarity with PF06737, which defines a domain that is likely to act as a transglycosylase. Alignment of the seed sequences of this Pfam domain with the four WY domains identified in *L. plantarum* proteins (supplementary Fig. S2) revealed that sequence similarity is best in the N-terminal half of the PF06737 domain and includes a universally conserved Glu residue that could act as a catalytic residue. Putative transglycosylases are of interest in the study of host–microbe interactions, as specific glycosylation patterns might influence the recognition of bacteria by their host (Chia *et al.*, 2001; Horn *et al.*, 1999). In addition to the WY domain, three of the four WY-domain-containing *L. plantarum* proteins also contain a LysM domain, which has peptidoglycan-binding properties (Bateman & Bycroft, 2000; Steen *et al.*, 2003). A search of the NCBI bacterial genome database revealed 17 WY-domain-containing proteins that are encoded by the genomes of lactobacilli, streptococci and enterococci (*Lactobacillus plantarum*, *Lactobacillus acidophilus*, *Lactobacillus johnsonii*, *Lactobacillus sakei*, *Streptococcus pneumoniae*, *Streptococcus agalactiae*, *Streptococcus thermophilus*, *Streptococcus mutans* and *Enterococcus faecalis*).

The cell-surface hydrolase (CSH) domain. A domain of approximately 250 amino acids was found in five *L. plantarum* proteins and was designated the cell-surface hydrolase (CSH) domain, based on its sequence similarity with a range of Pfam domains belonging to the clan of alpha/beta hydrolase-fold families. Multiple-sequence alignment of the *L. plantarum* CSH-containing proteins (Fig. S3) confirmed the presence and relative spacing of the conserved Ser, Asp and His residues, as well as the conserved GxSxG motif, that are characteristic for the alpha/beta hydrolase-fold family (Nardini & Dijkstra, 1999). These findings suggest that these *L. plantarum* proteins act as cell-surface hydrolases. Nevertheless, the exact function of the postulated hydrolases remains to be established since the alpha/beta hydrolase-fold family includes peptidases, lipases, esterases and dehalogenases. Another distinct set of eight homologous proteins of *L. plantarum* also have the characteristics of alpha/beta hydrolases, but differ in sequence alignment from the CSH domains, and are classified as members of the Pfam DUF915 family.

WxL domains. The domains WxL1 and WxL2 are found in surface proteins encoded in conserved gene clusters of

extracellular proteins. The names are based on two conserved WxL motifs present in both domains (Chaillou *et al.*, 2005; Kleerebezem *et al.*, 2003). The WxL family was divided into two subfamilies, WxL1 and WxL2, on the basis of sequence characteristics of the domains and the proteins containing the domain. The proteins encoded in the conserved clusters are hypothesized to form an extracellular enzyme complex functioning in carbon source acquisition (Siezen *et al.*, 2006). The *L. plantarum* genome encodes nine of these conserved clusters encoding a total of 13 and eight surface proteins with the WxL1 and WxL2 domains, respectively. The WxL domain and this conserved gene cluster are described in more detail by Siezen *et al.* (2006).

Low-complexity repeat regions

We identified 16 extracellular proteins encoded in the *L. plantarum* genome with low-complexity repeat regions of at least 15 residues (Table 3). Five proteins (lp_0197, lp_2486, lp_2796, lp_3059 and lp_3127) contain a low-complexity region of at least 50 residues separating a putative LPxTG-type peptidoglycan anchor from the rest of the protein, suggesting that the regions act as spacers that presumably position the functional domains of the protein outside the peptidoglycan layer. Four of these proteins contain predicted mucus-binding or collagen-binding domains (Fig. 3).

The *L. plantarum* genome encodes an unusually large protein (lp_1303a) of over 3300 residues with a Ser-Asp repeat of 1600 residues. In the ClfB protein of *S. aureus*, a long Ser-Asp repeat has been shown to play a role in the adherence of this bacterium to fibrinogen (Hartford *et al.*, 1997). The Ser-Asp region in ClfB is over 300 residues in size and separates the fibronectin-binding domain from its C-terminal peptidoglycan anchor; it probably functions as a spacer (Hartford *et al.*, 1997). In lp_1303a the repeat separates a C-terminal transmembrane anchor from a bacterial Ig-like domain, which is found in a variety of bacterial adhesion proteins (Pfam: PF02368).

Chromosomal location

Genes encoding putative extracellular proteins are not distributed evenly on the *L. plantarum* chromosome (Kleerebezem *et al.*, 2003). The genome contains two so-called lifestyle adaptation islands: regions thought to encode proteins playing a role in the adaptation of *L. plantarum* to its particular habitat and lifestyle. The most prominent lifestyle adaptation island (3080–3260 kb) encodes almost exclusively proteins for sugar transport, metabolism and regulation. The second region (2600–3000 kb) encodes a relatively high number of extracellular proteins, including six of the nine conserved cell-surface clusters (Siezen *et al.*, 2006). Although this second lifestyle island recognized in the *L. plantarum* genome does not display the exclusivity of encoded functions as is found in the sugar-related chromosomal island, it appears to be significantly enriched in genes encoding extracellular functions. This suggests that the adaptation of *L. plantarum* to its environment is focused on two important processes: (i) direct interaction with the environment through extracellularly exposed functions and (ii) adaptation of its carbohydrate metabolism to the available carbon sources.

Table 3. Low-complexity repeat regions

Only regions larger than 14 amino acids are shown.

ORF	Position	Unit size	Repeat unit*	Repeat number	Mismatches†	LPxTG anchor
lp_0197	58–82	1	S	25	0	Yes
lp_0197	798–911	3	PSE	38	0/2/2	No
lp_0374	86–285	4	KK..	50	0/17/./.	No
lp_0689	113–152	8	S..AA.S.	5	0/./1/1/./1/.	No
lp_1303a	1552–3148	2	DS	799	0/1	No
lp_2162	165–220	8	A.S..S.S	7	0/./2/./2/./1	No
lp_2486	736–787	4	TTAP	13	0/4/0/0	Yes
lp_2578	568–582	3	PG.	5	0/1/.	Yes
lp_2588	296–455	8	EKPG.TEP	20	0/3/1/2/./3/2/1	Yes
lp_2796	825–905	3	PE.	27	0/7/.	Yes
lp_2847	257–280	4	S.T.	6	0/./1/.	No
lp_2925	83–106	4	T.A.	6	0/./2/.	Yes
lp_2958	325–459	5	SSS..	27	0/2/9/./.	No
lp_3014	108–122	3	T.S	5	0/./1	No
lp_3059	1175–1252	6	P.QPE.	13	0/./2/2/2/.	Yes
lp_3093	165–188	4	A.S.	6	0/./2/.	No
lp_3127	1040–1097	2	P.	29	0	Yes

*Dots indicate variable positions. Only repeat units containing 50 % or less variable positions are shown.

†Numbers indicate the number of repeat units with a mismatch at specific positions. Positions are separated by a solidus.

Phylogenetic distribution of extracellular proteins

Homologues of the *L. plantarum* extracellular proteins were sought in all bacteria for which the complete genome sequence is available. The identified homologues were divided into sets with similar phylogenetic distributions (Table S18). The *L. plantarum* genome encodes 97 extracellular proteins which appear specific for this species, to date. Four of these proteins (lp_0197, lp_1643, lp_3059 and lp_3114) are predicted to be involved in adherence (Fig. 3), 21 are predicted to have an enzymic function, three are phage-related, while the remaining 69 proteins do not have a predicted function. In contrast, this analysis identified a well-defined set of extracellular proteins with a similar distribution in *Lactococcus lactis*, *Listeria* species, *Lactobacillus* species, *Enterococcus faecalis* and *Bacillus cereus*, all encoded in conserved gene clusters for extracellular proteins (Siezen *et al.*, 2006).

Some of the domains identified in extracellular proteins of *L. plantarum* are not found in *L. acidophilus* and *L. johnsonii*, both of which are found predominantly in the GI tract (Altermann *et al.*, 2005; Pridmore *et al.*, 2004). Striking examples are the collagen-binding domain (Pfam: PF05737) and the related Cna_B domain (Pfam: PF05738), present in three putative adhesion proteins of *L. plantarum*. Despite their absence in *L. acidophilus* and *L. johnsonii*, the phylogenetic distribution pattern of these domains clearly links them to the GI tract: both domains are found in proteins of bacteria linked to the GI tract, such as *Clostridium perfringens*, *Listeria monocytogenes* and *Enterococcus faecalis*. Apparently, different species of lactobacilli use different mechanisms for adherence to host tissue.

Comparative genome hybridization shows that 8 % of all *L. plantarum* WCFS1 genes are predicted to be missing in one or more other strains of *L. plantarum* (Molenaar *et al.*, 2005). For the proteins predicted to be extracellular, this is reduced to 5 % (Douwe Molenaar, personal communication). The relatively high level of conservation of this group of proteins among different strains supports their important role in the capacity of *L. plantarum* to adapt to and interact with its environment.

Concluding remarks

The extracellular proteins of a bacterium play an essential role in its interaction with its environment. We have performed an extensive bioinformatics analysis of all proteins of *L. plantarum* predicted to be extracellular. This analysis has led to the improvement of function prediction of 36 of these putative extracellular proteins compared to the annotation at the time of publication of the *L. plantarum* genome (Table S19). In addition, for a vast number of extracellular proteins of *L. plantarum*, information on domain composition and phylogenetic distribution could be added to the genome annotation database. The updated annotation and refined domain analyses facilitate and lead further efforts to functionally characterize these proteins, thereby contributing to a better understanding of the interaction of *L. plantarum* with its various habitats. All data on the extracellular proteins described in this report can be accessed through the *Lactobacillus plantarum* secretome database at http://www.cmbi.ru.nl/secretome.

ACKNOWLEDGEMENTS

We would like to thank Quinta Helmer for preliminary identification and annotation of extracellular proteins and domains. This work was supported by a grant from the Netherlands Organization for Scientific Research (NWO-BMI project 050.50.206).

REFERENCES

Adlerberth, I., Ahrne, S., Johansson, M. L., Molin, G., Hanson, L. A. & Wold, A. E. (1996). A mannose-specific adherence mechanism in *Lactobacillus plantarum* conferring binding to the human colonic cell line HT-29. *Appl Environ Microbiol* **62**, 2244–2251.

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.

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.

Bateman, A. & Bycroft, M. (2000). The structure of a LysM domain from *E. coli* membrane-bound lytic murein transglycosylase D (MltD). *J Mol Biol* **299**, 1113–1119.

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

Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783–795.

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.

Boekhorst, J., Helmer, Q., Kleerebezem, M. & Siezen, R. J. (2006). Comparative analysis of proteins with a mucus-binding domain found exclusively in lactic acid bacteria. *Microbiology* **152**, 273–280.

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

Brendel, V., Bucher, P., Nourbakhsh, I. R., Blaisdell, B. E. & Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci U S A* **89**, 2002–2006.

Buck, B. L., Altermann, E., Svingerud, T. & Klaenhammer, T. R. (2005). Functional analysis of putative adhesion factors in *Lactobacillus acidophilus* NCFM. *Appl Environ Microbiol* **71**, 8344–8351.

Cabanes, D., Dehoux, P., Dussurget, O., Frangeul, L. & Cossart, P. (2002). Surface proteins and the pathogenic potential of *Listeria monocytogenes*. *Trends Microbiol* **10**, 238–245.

Chaillou, S., Champomier-Verges, M. C., Cornet, M. & 8 other authors (2005). The complete genome sequence of the meat-borne lactic acid bacterium *Lactobacillus sakei* 23K. *Nat Biotechnol* **23**, 1527–1533.

Chia, J. S., Chang, L. Y., Shun, C. T., Chang, Y. Y., Tsay, Y. G. & Chen, J. Y. (2001). A 60-kilodalton immunodominant glycoprotein

is essential for cell wall integrity and the maintenance of cell shape in *Streptococcus mutans*. *Infect Immun* **69**, 6987–6998.

Christie, J., McNab, R. & Jenkinson, H. F. (2002). Expression of fibronectin-binding protein FbpA modulates adhesion in *Streptococcus gordonii*. *Microbiology* **148**, 1615–1625.

de Vries, M., Vaughan, E. E., Kleerebezem, M. & de Vos, W. M. (2005). *Lactobacillus plantarum* – survival, functional and potential probiotic properties in the human intestinal tract. *Int Dairy J* **16**, 1018–1028.

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

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

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

Gaillard, J. L., Berche, P., Frehel, C., Gouin, E. & Cossart, P. (1991). Entry of *L. monocytogenes* into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. *Cell* **65**, 1127–1141.

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.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696–704.

Hammerschmidt, S. (2006). Adherence molecules of pathogenic pneumococci. *Curr Opin Microbiol* **9**, 12–20.

Hartford, O., Francois, P., Vaudaux, P. & Foster, T. J. (1997). The dipeptide repeat region of the fibrinogen-binding protein (clumping factor) is required for functional expression of the fibrinogen-binding domain on the *Staphylococcus aureus* cell surface. *Mol Microbiol* **25**, 1065–1076.

Horn, C., Namane, A., Pescher, P., Riviere, M., Romain, F., Puzo, G., Barzu, O. & Marchal, G. (1999). Decreased capacity of recombinant 45/47-kDa molecules (Apa) of *Mycobacterium tuberculosis* to stimulate T lymphocyte responses related to changes in their mannosylation pattern. *J Biol Chem* **274**, 32023–32030.

Jankovic, I., Ventura, M., Meylan, V., Rouvet, M., Elli, M. & Zink, R. (2003). Contribution of aggregation-promoting factor to maintenance of cell shape in *Lactobacillus gasseri* 4B2. *J Bacteriol* **185**, 3288–3296.

Kawai, R., Igarashi, K. & Samejima, M. (2006). Gene cloning and heterologous expression of glycoside hydrolase family 55 beta-1,3-glucanase from the basidiomycete *Phanerochaete chrysosporium*. *Biotechnol Lett* **28**, 365–371.

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.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 567–580.

Marco, M. L., Pavan, S. & Kleerebezem, M. (2006). Towards understanding molecular modes of probiotic action. *Curr Opin Biotechnol* **17**, 204–210.

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

Nardini, M. & Dijkstra, B. W. (1999). Alpha/beta hydrolase fold enzymes: the family keeps growing. *Curr Opin Struct Biol* **9**, 732–737.

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.

Ni Eidhin, D., Perkins, S., Francois, P., Vaudaux, P., Hook, M. & Foster, T. J. (1998). Clumping factor B (ClfB), a new surface-located fibrinogen-binding adhesin of *Staphylococcus aureus*. *Mol Microbiol* **30**, 245–257.

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.

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

Siezen, R. J., Boekhorst, J., Muscariello, L., Molenaar, D., Renckens, B. & Kleerebezem, M. (2006). *Lactobacillus plantarum* gene clusters encoding putative cell-surface protein complexes for carbohydrate utilization are conserved in specific gram-positive bacteria. *BMC Genomics* **7**, 126.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol* **147**, 195–197.

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

Steen, A., Buist, G., Leenhouts, K. J., El Khattabi, M., Grijpstra, F., Zomer, A. L., Venema, G., Kuipers, O. P. & Kok, J. (2003). Cell wall attachment of a widely distributed peptidoglycan binding domain is hindered by cell wall constituents. *J Biol Chem* **278**, 23874–23881.

Sutcliffe, I. C. & Harrington, D. J. (2002). Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. *Microbiology* **148**, 2065–2077.

Sutcliffe, I. C. & Russell, R. R. (1995). Lipoproteins of gram-positive bacteria. *J Bacteriol* **177**, 1123–1128.

Symersky, J., Patti, J. M., Carson, M. & 8 other authors (1997). Structure of the collagen-binding domain from a *Staphylococcus aureus* adhesin. *Nat Struct Biol* **4**, 833–838.

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.

Ventura, M., Jankovic, I., Walker, D. C., Pridmore, R. D. & Zink, R. (2002). Identification and characterization of novel surface proteins in *Lactobacillus johnsonii* and *Lactobacillus gasseri*. *Appl Environ Microbiol* **68**, 6172–6181.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. & Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33**, D433–D437.

Wels, M., Francke, C., Kerkhoven, R., Kleerebezem, M. & Siezen, R. J. (2006). Predicting *cis*-acting elements of *Lactobacillus plantarum* by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res* **34**, 1947–1958.

Yuen, L., Dionne, J., Arif, B. & Richardson, C. (1990). Identification and sequencing of the spheroidin gene of *Choristoneura biennis* entomopoxvirus. *Virology* **175**, 427–433.

Zhang, J. R., Mostov, K. E., Lamm, M. E., Nanno, M., Shimida, S., Ohwaki, M. & Tuomanen, E. (2000). The polymeric immunoglobulin receptor translocates pneumococci across human nasopharyngeal epithelial cells. *Cell* **102**, 827–837.