JOURNAL OF BACTERIOLOGY, July 2005, p. 4928–4934 Vol. 187, No. 14
0021-9193/05/$08.00+0 doi:10.1128/JB.187.14.4928–4934.2005

Genome-Wide Detection and Analysis of Cell Wall-Bound Proteins with LPxTG-Like Sorting Motifs

Jos Boekhorst,[1]* Mark W. H. J. de Been,[1,2] Michiel Kleerebezem,[2,3] and Roland J. Siezen[1,2,3]

Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen, 6525ED Nijmegen, The Netherlands[1]; Wageningen Centre for Food Sciences, Wageningen, The Netherlands[2]; and NIZO Food Research, Ede, The Netherlands[3]

Received 19 January 2005/Accepted 17 April 2005

Surface proteins of gram-positive bacteria often play a role in adherence of the bacteria to host tissue and are frequently required for virulence. A specific subgroup of extracellular proteins contains the cell wall-sorting motif LPxTG, which is the target for cleavage and covalent coupling to the peptidoglycan by enzymes called sortases. A comprehensive set of putative sortase substrates was identified by in silico analysis of 199 completely sequenced prokaryote genomes. A combination of detection methods was used, including secondary structure prediction, pattern recognition, sequence homology, and genome context information. With the hframe algorithm, putative substrates were identified that could not be detected by other methods due to errors in open reading frame calling, frameshifts, or sequencing errors. In total, 732 putative sortase substrates encoded in 49 prokaryote genomes were identified. We found striking species-specific variation for the LPxTG motif. A hidden Markov model (HMM) based on putative sortase substrates was created, which was subsequently used for the automatic detection of sortase substrates in recently completed genomes. A database was constructed, LPxTG-DB (http://bamics3.cmbi.kun.nl/sortase_substrates), containing for each genome a list of putative sortase substrates, sequence information of these substrates, the organism-specific HMMs based on the consensus sequence of the sortase recognition motif, and a graphic representation of this consensus.

The cell wall of gram-positive bacteria consists of many different types of macromolecules. It includes covalently and noncovalently linked proteins, carbohydrates, polysaccharides, and teichoic acids, embedded in a peptidoglycan matrix (24). The peptidoglycan matrix protects the cell against both mechanical and osmotic lysis and plays an important role in the interaction of the cell with its surroundings (15). Bacterial host infections, for instance, are often mediated by many of the covalently linked surface proteins (14, 20).

One class of covalently bound surface proteins is characterized by a cell wall-sorting motif called LPxTG (based on the main conserved residues). The motif is located at the C terminus of the protein, followed by a stretch of hydrophobic residues and a number of positively charged amino acids (13, 29). The hydrophobic domain and the charged tail probably keep the protein from being secreted into the medium, thereby allowing recognition of the LPxTG motif by a membrane-associated transpeptidase called sortase. Sortase cleaves the LPxTG motif between the T and G residues and covalently attaches the threonine carboxyl group to the peptidoglycan (23).

Not all proteins that have been experimentally verified to be sortase substrates contain a cell wall-sorting motif that fits the pattern LPxTG. The sortase SrtB from *Staphylococcus aureus* recognizes the motif NPQTN (21), and Bierne and coworkers showed that a protein with an NAKTN motif is attached to the cell wall of *Listeria monocytogenes* by a sortase-like enzyme (8). Recently a protein with the strongly deviating QVPTGV motif was discovered to be a sortase substrate (5).

Many bacterial genomes encode more than one sortase (28), and five distinct subfamilies can be distinguished among these transpeptidases (10). It has been suggested that it is possible to predict the specificity of a sortase for a group of substrates based on the amino acid sequence of the sortase, the cell wall-sorting signal of potential substrates, and the relative positioning of genes encoding sortases and substrates on the bacterial chromosome (10). Genome context in particular seems a strong indicator of functional relationship, as sortases and their substrates are often encoded in gene clusters on bacterial chromosomes.

In this study, a comprehensive set of putative sortase substrates was identified by in silico analysis of 199 sequenced bacterial genomes.

Since the sortase recognition sequence LPxTG itself is very short, searching only for this motif (and its variants) will lead to many incorrect predictions which, based on other characteristics of these hits such as predicted number of transmembrane helices and predicted protein function, are probably not sortase substrates. Therefore, we have applied a combination of methods, including secondary structure prediction, pattern detection, genome context, and homolog detection, to reduce the number of incorrect predictions. Some bacteria preferentially encode sortase substrates that contain target sequences deviating slightly from the canonical LPxTG motif. The predicted sortase substrates of *Lactobacillus plantarum*, for example, contain an LPQTxE motif instead of an LPxTG motif (17). Because of this variation, optimization of the sequence pattern used for the detection of sortase substrates for a specific bacterium increases the sensitivity and selectivity of the analysis

* Corresponding author. Mailing address: Center for Molecular and Biomolecular Informatics, Radboud University Nijmegen, 6525ED Nijmegen, The Netherlands. Phone: 31 24 3653398. Fax: 31 24 3652977. E-mail: J.Boekhorst@cmbi.ru.nl.

(16). We have applied species-specific hidden Markov models (HMMs) to identify putative sortase substrates and have determined the extent and nature of the species-specific variation for the LPxTG motif. Use of the hframe algorithm allowed us to detect putative sortase substrates on the DNA level that were not detected by the other methods, for example, due to errors in open reading frame calling.

MATERIALS AND METHODS

Sequence information. Genome sequence information was obtained from the National Center for Biotechnology Information (NCBI) bacterial genome database (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). All 154 complete bacterial genomes present in this database on 21 February 2004 were used for developing the search routines. The LPxTG-HMM generated in this way was subsequently used to search in all 199 bacterial genomes present on 24 November 2004.

Sequence analysis. Sequence similarity was detected with BLAST (1), while multiple sequence alignments were made with T-Coffee (27). Transmembrane helices were predicted with TMHMM 2 (18), and signal peptides were predicted with SignalP 2.0 (26).

The HMMER package (12) was used to construct HMMs based on these alignments and to scan protein sequences with HMMs. Pattern recognition analysis was performed with FindPatterns (32). Conserved sequence patterns were identified with MEME and MAST (3, 4). The hframe algorithm provided by Paracel was used to scan translated nucleotide sequences with protein-based HMMs.

Identification of sortase enzymes. Two HMMs from the Pfam database (7) were used to detect sortases: the sortase A HMM (PF04203, sortase) and the sortase B HMM (PF07170, sortase_B). All protein sequences were scanned with these HMMs, and all proteins with an E-score below 1e-05 were considered putative sortases. A search of the NCBI bacterial genome database for proteins annotated as sortases did not yield any additional hits.

Identification of sortase substrates. The identification of putative sortase substrates was performed as described below and is depicted in Fig. 1 (an in-depth description of these methods can be found at http://bamics3.cmbi.kun.nl/sortase_substrates/supplementary).

For each organism, two methods were used to compile an initial set of putative sortase substrates. The first method involved using the program FindPatterns to identify putative sortase substrates by scanning all protein sequences of the bacterium with a regular expression describing the sortase cleavage site, the C-terminal helix, and the positive charge following this helix (the "tripartite pattern") (16). Based on this initial set of putative substrates identified, a species-specific HMM was created which was subsequently used to identify additional substrates in the corresponding genome.

The second method involved the use of MEME and MAST to predict sortase substrates. The last 60 amino acids of all proteins containing a signal peptide were used as input for a MEME motif search. From the resulting list of motifs, the pattern with the highest resemblance to the C terminus of known sortase substrates was used in a genome-wide MAST search. For each organism, no more than one pattern was found that fit the characteristics of a cell wall-sorting signal. The results of the FindPatterns-HMM and MEME-MAST methods were combined to create an improved set of predicted sortase substrates.

Additional substrates were found (i) by identifying proteins homologous to the putative sortase substrates of the improved set and (ii) by checking all proteins in gene clusters containing at least one sortase substrate or sortase enzyme. Then, on the basis of the resulting complete set, a new HMM was created which was used to rescan all protein sequences and to scan all chromosomal DNA sequences using the hframe algorithm, resulting in a final set of putative sortase substrates.

RESULTS AND DISCUSSION

Identification of sortase substrates. We extracted genome sequence information of 154 bacteria from the NCBI genome database and searched for sortases and their substrates, as described in Fig. 1. The results are summarized in Table 1. We predicted a total of 568 sortase substrates in 39 of these genomes, of which 531 were identified by the FindPatterns-HMM method and 495 by the MEME-MAST method. Combination of the output of these two methods led to the prediction of 533 candidate sortase substrates. The use of MEME did not significantly increase the number of putative sortase substrates identified; only two additional substrates were found, probably because the MEME method searches for sequence patterns that are significantly overrepresented in a set of protein sequences. It will therefore not find a pattern present in only a couple of proteins in an entire genome.

Additional searches using homology and genome context search yielded 12 and 6 additional putative substrates, respectively. Finally, the analysis of chromosomal DNA using the hframe algorithm led to the identification of an additional 13 predicted substrates, the majority of which had been undetected by the other methods, due to errors in the identification of protein-encoding genes in the bacterial genome sequences (see Table 3).

For each genome studied, a list of putative sortase substrates, sequence information of these substrates, the HMM based on the consensus sortase recognition signal of this bacterium, and a graphic representation of this consensus sequence can be found in the LPxTG-DB database at http://bamics3.cmbi.kun.nl/sortase_substrates.

Inspection of putative sortase substrate sequences showed that many proteins are detected with one or more mismatches in the LPxTG-like motif. Nevertheless, these proteins all met the criteria for sortase substrates as outlined in Materials and Methods. We evaluated the sensitivity of our method by searching the literature for proteins that were experimentally verified to be attached to the bacterial cell wall in a sortase-dependent manner. All of the 24 proteins for which we found experimental verification (5, 6, 8, 9, 19, 21, 25) were present in our data set of predicted sortase substrates, including those with highly deviating LPxTG-like motifs, illustrating the high sensitivity of our methods. These substrates are listed at http://bamics3.cmbi.kun.nl/sortase_substrates/supplementary.

Newly identified sortase substrates. The first set of putative sortase substrates we found by the initial FindPatterns-HMM and MEME-MAST methods was similar to the set of putative substrates that others have identified using methods very similar to the FindPatterns-HMM method (10). However, in the same set of genomes we found 65 additional putative sortase substrates (11% more) that were not identified by their methods. Most of the additional 65 putative substrates were identified with the help of homology, genome context, and the use of the hframe algorithm. Manual inspection showed that the main reasons why these additional substrates were not detected by the FindPatterns-HMM and MEME-MAST methods were either (i) the deviation of some organism-specific sortase cleavage motifs from the generic LPxTG motif, (ii) the lack of a signal peptide (caused, for example, by the incorrect prediction of translation starts), or (iii) substrates not previously being recognized as protein-encoding genes.

Eight genomes contain at least one predicted sortase gene, while no sortase substrates were predicted by either the FindPatterns-HMM or MEME-MAST methods (Tables 1 and 2). In six of these genomes, one or two sortase substrates could be predicted by one of our other methods. One of these proteins, the single putative sortase substrate of *Bradyrhizobium japonicum*, had not been previously identified. The other proteins were already classified as putative sortase sub-

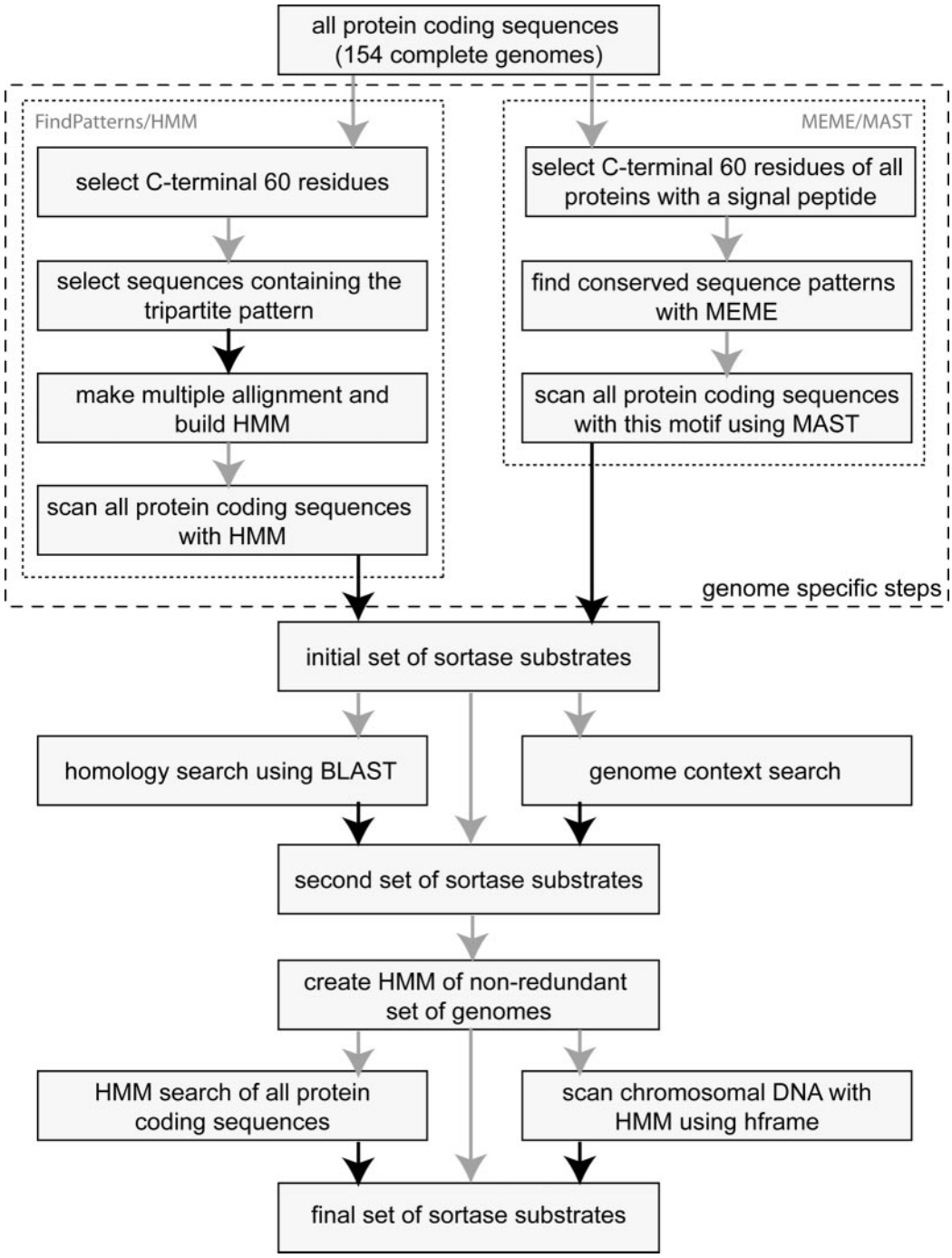

FIG. 1. Detecting sortase substrates. The steps shown in the dashed rectangle were carried out for each of the 154 genomes individually. Gray arrows indicate that all proteins meeting the selection criteria described in the box were taken to the next step. Black arrows indicate that the proteins had to meet additional criteria, as follows. (i) Proteins should have a transmembrane helix following the sortase recognition motif LPxTG. (ii) This helix should be followed by positively charged amino acid residues. (iii) Proteins should have three or fewer transmembrane helices in their complete precursor sequence. (iv) Proteins should not have a predicted function indicating intracellular localization.

strates by Interpro (22). In the two genomes without predicted sortase substrates (*Methanobacterium thermoautotrophicum* and *Corynebacterium glutamicum*), the role of the sortase-like transpeptidases remains unclear.

Using the final HMM to identify sortase substrates. We created a HMM, coined LPxTG-HMM, based on a multiple sequence alignment of the C-terminal 60 residues of all putative sortase substrates identified by the FindPatterns-HMM, MEME-MAST, homology, and context searches. To determine the value of this LPxTG-HMM as a tool for quickly identifying putative sortase substrates in large data sets, we used it to scan the C-terminal fragments of all of the proteins

TABLE 1. Predicted sortase substrates in original set of 154 genomes

Species	No. of sortase substrates[a]							No. of sortases
	FP/HMM	MEME/MAST	Additional hits				Total	
			BLAST	Context	LPxTG-HMM	hframe		
Actinobacteria (high-G+C gram-positive bacteria)								
Corynebacterium diphtheriae NCTC13129	16	16	0	0	0	1	17	6
Corynebacterium efficiens YS-314	8	0	0	0	0	0	8	5
C. glutamicum ATCC 13032	0	0	0	0	0	0	0	1
S. avermitilis MA-4680	14	13	1	0	0	0	16	9
S. coelicolor A3(2)	15	15	0	0	0	1	17	7
Tropheryma whipplei TW0827	0	0	0	0	1	0	1	1
Tropheryma whipplei Twist	0	0	0	0	1	0	1	1
Bifidobacterium longum NCC2705	16	16	0	0	0	1	17	3
Firmicutes (gram-positive bacteria)								
B. anthracis A2012	10	8	2	0	0	0	12	3
B. anthracis Ames	9	7	1	0	0	2	12	3
B. cereus ATCC14579	14	12	1	0	0	1	16	5
B. cereus ATCC10987	14	14	1	0	0	2	17	6
B. halodurans C-125	9	8	0	0	0	0	9	6
B. subtilis subsp. *subtilis* 168	0	0	2	0	0	0	2	2
Clostridium acetobutylicum ATCC 824	2	0	0	0	0	0	2	1
Clostridium perfringens 13	13	12	0	0	0	0	13	5
Clostridium tetani E88	3	0	0	0	0	0	3	1
Enterococcus faecalis V583	33	33	0	0	1	1	35	3
L. johnsonii NCC533	16	16	0	0	0	0	16	2
L. plantarum WCFS1	27	26	0	0	0	0	27	1
Lactococcus lactis subsp. *lactis* IL1403	11	7	0	0	0	1	12	2
L. innocua CLIP 11262	35	34	1	0	0	0	36	2
L. monocytogenes EGD-e	43	40	0	0	0	0	43	2
Oceanobacillus iheyensis HTE831	3	0	0	0	0	0	3	3
S. aureus Mu50	19	19	1	0	0	0	20	2
S. aureus MW2	20	20	1	0	0	0	21	2
S. aureus N315	18	18	1	0	0	1	20	2
Staphylococcus epidermidis ATCC_12228	10	10	0	0	0	0	10	2
Streptococcus agalactiae 2603	25	25	0	0	0	0	25	6
S. agalactiae NEM316	35	35	0	0	0	0	35	5
Streptococcus mutans UA159	6	6	0	0	0	0	6	1
S. pneumoniae R6	14	14	0	0	0	0	14	1
S. pneumoniae TIGR4	15	15	0	0	0	1	16	4
Streptococcus pyogenes M1GAS	14	14	0	1	0	0	15	3
S. pyogenes MGAS315	15	15	0	1	0	0	16	2
S. pyogenes MGAS8232	14	13	0	1	0	1	16	2
S. pyogenes SSI-1	15	14	0	1	0	0	16	2
Proteobacteria (gram-negative bacteria)								
B. japonicum USDA 110	0	0	0	1	0	0	1	1
Shewanella oneidensis MR-1	0	0	0	1	0	0	1	1
Archaea								
M. thermoautotrophicum Delta H	0	0	0	0	0	0	0	2
Methanopyrus kandleri AV19	0	0	0	0	1	0	1	1
Total	531	495	12	6	4	13	568	119

[a] FP, FindPatterns; additional hits, putative sortase substrates identified in addition to the set of putative substrates found by the FindPatterns/HMM and MEME/MAST methods.

encoded by the 41 prokaryote genomes with at least one sortase. The LPxTG-HMM identified 553 of the 564 proteins detected by one of the other methods. Of the 15 proteins not detected with the LPxTG-HMM, 13 were detected with hframe. These putative substrates could not be detected by the HMM because they had not previously been identified as protein-coding sequences (CDSs). When these proteins were not taken into account, the LPxTG-HMM by itself identified over 99% of the total number of putative sortase substrates identified by our combination of methods. One of the two missed proteins was a putative sortase substrate from *Streptomyces avermitilis* with an LAETG cleavage site, which actually fits the organism-specific cleavage consensus of *S. avermitilis* (i.e., LAxTG) quite well. However, this protein scored too low against the final HMM because of the alanine residue at the second position of the recognition site in combination with a

TABLE 2. Predicted sortase substrates in 45 recently sequenced genomes

Firmicute (gram-positive bacteria) species	No. of sortase substrates[a]							No. of sortases
	FP/HMM	MEME/MAST	Additional hits				Total	
			BLAST	Context	LPxTG-HMM	hframe		
B. anthracis Ames 0581	—[b]	—	—	—	11	2	13	3
B. anthracis Sterne	—	—	—	—	12	0	12	3
B. cereus ZK	—	—	—	—	15	0	15	3
Bacillus licheniformis ATCC 14580	—	—	—	—	4	0	4	3
B. licheniformis DSM 13	—	—	—	—	4	0	4	3
Bacillus thuringiensis konkukian	—	—	—	—	12	0	12	3
L. monocytogenes 4b F2365	—	—	—	—	47	1	48	2
S. aureus subsp. *aureus* MRSA252	—	—	—	—	16	2	18	2
S. aureus subsp. *aureus* MSSA476	—	—	—	—	20	2	22	2
S. pyogenes MGAS10394	—	—	—	—	16	0	16	2
Total					157	7	164	26

[a] For abbreviations, see Table 1, footnote *a*.

relatively small positive charge at the C terminus of the protein. This illustrates the value of using a species-specific HMM for the detection of putative sortase substrates in organisms with a consensus sortase cleavage site that deviates from the generic LPxTG consensus. The other protein missed by the LPxTG-HMM was a putative sortase substrate with an NSKTA cleavage signal in *Bacillus cereus* ATCC 14579. A protein *of L. monocytogenes* orthologous to this *B. cereus* protein has been experimentally proven to be a sortase substrate (8); in other strains of *B. cereus*, *Bacillus anthracis*, and *Bacillus halodurans*, putative sortase substrates with this cleavage signal were detected.

With a bit-score threshold of 5, the LPxTG-HMM predicted 34 potential sortase substrates not identified by any of the other methods. Only four of these fulfilled the criteria of sortase substrates as described in Materials and Methods and unpublished data and hence were added to Tables 1 and 2. The other 30 proteins (5% of the total number of hits) did not meet these criteria, for example, due to the presence of too many predicted transmembrane helices. The bit score threshold of 5 was determined empirically: a higher threshold causes many proteins fitting the criteria for sortase substrates as outlined in Materials and Methods to be missed, while a lower threshold of 4 leads to the inclusion of many proteins with a C-terminal membrane helix, followed by positively charged residues, but without an LPxTG-like motif.

As mentioned earlier, application of the hframe algorithm revealed 13 additional genes encoding putative substrates (Table 3). Furthermore, the hframe algorithm identified another six sequences with all of the characteristics of sortase substrates, but for which no correct translation start could be identified without introducing a frameshift or removing an internal stop codon. In some cases, the introduction of a frameshift or the removal of a stop codon would merge a novel CDS encoding a putative sortase substrate (i.e., not previously recognized as a CDS) with a CDS already identified on the chromosome. It remains to be established whether these six additional sequences represent pseudogenes or sequencing errors.

Compared to the gram-positive anchor HMMs and suggested thresholds of the Pfam (7) and TIGRFAM (http://www.tigr.org/TIGRFAMs/) databases, LPxTG-HMM detects many more putative sortase substrates. Although the LPxTG-HMM slightly overpredicted the number of sortase substrates, the incorrectly identified substrates (i.e., proteins not fitting the criteria for sortase substrates as outlined in the methods section) were easily filtered out by application of the simple additional criteria mentioned in Materials and Methods. Fur-

TABLE 3. Newly identified CDSs[a]

Organism	Direction	Start	Stop	Reason
Streptomyces coelicolor A3(2)	−	5367642	5367981	not defined
Streptococcus pyogenes MGAS8232	+	854197	854893	not defined
Streptococcus pneumoniae TIGR4	+	341382	346685	not defined
Staphylococcus aureus N315	−	2559703	2562486	frameshift
Lactococcus lactis subsp. *lactis* IL1403	−	1627879	1628064	frameshift
Enterococcus faecalis V583	+	556642	558252	not defined
Corynebacterium diphtheriae NCTC13129	−	2264577	2264876	not defined
Bifidobacterium longum NCC2705	+	190516	190776	not defined
Bacillus cereus ATCC 10987	+	1742239	1743951	not defined
Bacillus cereus ATCC 10987	+	1738923	1742249	not defined
Bacillus cereus ATCC 14579	−	4085606	4087576	not defined
Bacillus anthracis Ames	+	5092743	5095385	not defined
Bacillus anthracis Ames	−	4686784	4691070	not defined

[a] The column "Reason" provides an explanation for the CDS not being identified as CDS previously. Not defined, ORF present but no CDS defined; frameshift, the new CDS is adjacent to an already defined CDS but not part of this CDS because of a frameshift.

thermore, LPxTG-HMM outperformed the other methods in the detection of sortase substrates with a sortase recognition signal deviating from the consensus signal. For example, only 2 of the 17 sortase substrates of *Streptomyces coelicolor* are detected by the Pfam and TIGRFAM HMMs.

To determine whether or not cell wall-sorting-like signals are only present in the C termini of proteins, we scanned the complete sequences of all of the proteins taken from the NCBI bacterial genome database with the LPxTG-HMM. We identified only three proteins with a putative cell wall-sorting signal at a position other than the C terminus: two proteins with orthologs in *Streptococcus pneumoniae* R6 and *S. pneumoniae* TIGR4 and one protein with orthologs in *L. monocytogenes* EGD-e, *L. monocytogenes* 4b F2365, and *Listeria innocua*. The presence of orthologs in different strains indicates that these proteins are not the results of a sequencing anomaly (e.g., a frameshift caused by a sequencing error, leading to the fusion of two CDSs). All three proteins contained an N-terminal signal peptide, and the predicted function of the two pneumococcal proteins was consistent with an extracellular localization: one of the proteins was predicted to be a zinc metalloprotease, and the other was predicted to be an immunoglobulin A1 protease. The unusual position of the LPxTG motif in these sequences could be the result of a gene fusion event. The N-terminal parts of the three proteins did not have significant sequence homology to any sequence in the UniProt protein database (2).

Signal peptides. Each protein that is destined to become attached to the peptidoglycan via the LPxTG anchor should also have an N-terminal signal peptide with consensus cleavage motif AxA ↓ A (30, 31) for initial translocation of the protein across the cell membrane. Nevertheless, of our final list of 568 putative sortase substrates identified, 56 did not appear to have a signal peptide (as predicted by SignalP). However, upon closer inspection we were able to identify an N-terminal signal peptide for 43 of them (http://bamics3.cmbi.kun.nl/sortase_substrates/supplementary). In 25 cases, this required the selection of a different start codon than the one specified by the NCBI genome annotation; in 5 cases, this required the removal of a stop codon; and in 13 cases, it required the introduction of a frameshift. To determine whether or not such a stop codon or frameshift could be the result of a sequencing error would require access to the trace files of the sequencing projects. The gene identifiers and suggested changes to the CDSs for the 56 predicted sortase substrates without a signal peptide are shown at http://bamics3.cmbi.kun.nl/sortase_substrates/supplementary.

Species-specific anchoring motifs. Closely related organisms have similar sortase recognition consensus sequences, leading to similar HMMs. For instance, the organism-specific HMMs of *B. anthracis* Ames and *B. cereus* ATCC 10987 detect the same set of 10 putative sortase substrates in the *B. anthracis* genome. As expected, HMMs from less-similar organisms have less overlap; when the HMM based on the putative sortase substrates of *S. coelicolor* is used to scan the *B. anthracis* genome, only two putative substrates were recognized.

A graphic representation of the species-specific LPxTG consensus of every bacterium with two or more predicted sortase substrates can be found in our LPxTG-DB database (http://bamics3.cmbi.kun.nl/sortase_substrates). In some organisms, many putative sortase substrates have a cleavage motif that is highly conserved, but which deviates significantly from the generic LPxTG consensus and the motifs found in other organisms. Examples of such organisms and the frequency with

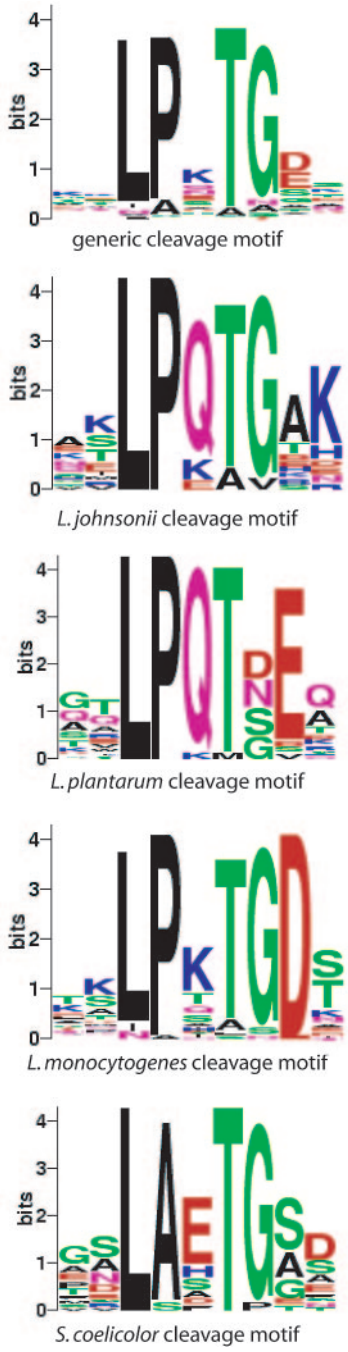

FIG. 2. Organism-specific cleavage motifs. The consensus sortase cleavage sites of *L. plantarum* (LPQTxE, found in 23 of 27 predicted sortase substrates), *Lactobacillus johnsonii (*LPQTG, found in 12 of 16 substrates), *L. monocytogenes* (LPxTGD, found in 33 of 42 substrates), and *S. coelicolor* (LAxTG, found in 15 of 17 substrates) are organism-specific variations on the generic LPxTG consensus. The overall height of each stack indicates the sequence conservation at that position (measured in bits), whereas the height of symbols within the stack reflects the relative frequency of the corresponding amino acid at that position. The Weblogo software (11) was used to visualize the motifs.

which specific motifs are found in these organisms are shown in Fig. 2. The fact that these motifs are highly conserved suggests that these sortase substrates are species specific and also implies they have not been acquired through horizontal gene transfer or are rapidly optimized due to selective pressure.

Function of sortase substrates. Of the 568 putative sortase substrates identified by us, 67% do not have any predicted function, 15% are predicted to have an enzymatic function, and 10% are predicted to have a binding function (e.g., collagen-binding protein). The predicted functions, as in the original annotation at NCBI, can be found in the LPxTG-DB database. Large differences in the methods used for the functional annotation of the different genomes make it difficult to compare sortase substrate functions between genomes. A better approach is to predict the function of putative sortase substrates by determining their domain composition with the Pfam (7) and Interpro (22) databases.

Searching in new genomes. Finally, we used the LPxTG-HMM to identify putative sortase substrates in the 45 new genomes that were made public after the date on which we took our original set of genomes from GenBank. For 10 of these 45 additional genomes, all from gram-positive bacteria, we predicted a total of 164 sortase substrates (Table 2), 7 of which had not been identified as CDSs in the GenBank annotation. The other 35 genomes did not encode any putative sortase substrates or sortases. The results of this analysis can also be found in our database of sortase substrates, LPxTG-DB.

Concluding remarks. We developed an HMM which quickly and reliably recognizes the putative sortase substrates in any sequenced genome. Although the model does not incorporate explicitly all of the information available, when used together with the hframe algorithm it recovers >99% of the putative substrates detected by several other methods combined. When the combination of methods we have described in this research is used, an average of 11% additional putative sortase substrates can be identified compared to previously used methods.

Our sortase-substrate website contains information on the species-specific sortase recognition sites identified, the LPxTG-HMM, and brief instructions on its use.

ACKNOWLEDGMENTS

We thank Christof Francke for careful reading of the manuscript.

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

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, and L. S. Yeh.** 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. **32:**D115–D119.
3. **Bailey, T. L., and C. Elkan.** 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. **2:**28–36.
4. **Bailey, T. L., and M. Gribskov.** 1998. Combining evidence using p-values: application to sequence homology searches. Bioinformatics **14:**48–54.
5. **Barnett, T. C., A. R. Patel, and J. R. Scott.** 2004. A novel sortase, SrtC2, from *Streptococcus pyogenes* anchors a surface protein containing a QVPTGV motif to the cell wall. J. Bacteriol. **186:**5865–5875.
6. **Barnett, T. C., and J. R. Scott.** 2002. Differential recognition of surface proteins in *Streptococcus pyogenes* by two sortase gene homologs. J. Bacteriol. **184:**2181–2191.
7. **Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy.** 2004. The Pfam protein families database. Nucleic Acids Res **32**(database issue)**:**D138–D141.
8. **Bierne, H., C. Garandeau, M. G. Pucciarelli, C. Sabet, S. Newton, F. Garcia-del Portillo, P. Cossart, and A. Charbit.** 2004. Sortase B, a new class of sortase in *Listeria monocytogenes*. J. Bacteriol. **186:**1972–1982.
9. **Bierne, H., S. K. Mazmanian, M. Trost, M. G. Pucciarelli, G. Liu, P. Dehoux, L. Jansch, F. Garcia-del Portillo, O. Schneewind, and P. Cossart.** 2002. Inactivation of the srtA gene in Listeria monocytogenes inhibits anchoring of surface proteins and affects virulence. Mol. Microbiol. **43:**869–881.
10. **Comfort, D., and R. T. Clubb.** 2004. A comparative genome analysis identifies distinct sorting pathways in gram-positive bacteria. Infect. Immun. **72:**2710–2722.
11. **Crooks, G. E., G. Hon, J. M. Chandonia, and S. E. Brenner.** 2004. WebLogo: a sequence logo generator. Genome Res. **14:**1188–1190.
12. **Durbin, R., S. R. Eddy, A. Krogh, and G. Mitchison.** 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, United Kingdom.
13. **Fischetti, V. A., V. Pancholi, and O. Schneewind.** 1990. Conservation of a hexapeptide sequence in the anchor region of surface proteins from gram-positive cocci. Mol. Microbiol. **4:**1603–1605.
14. **Foster, T. J., and M. Hook.** 1998. Surface protein adhesins of Staphylococcus aureus. Trends Microbiol. **6:**484–488.
15. **Ghuysen, J.** 1994. Bacterial cell wall. Elsevier Science, Amsterdam, The Netherlands.
16. **Janulczyk, R., and M. Rasmussen.** 2001. Improved pattern for genome-based screening identifies novel cell wall-attached proteins in gram-positive bacteria. Infect. Immun. **69:**4019–4026.
17. **Kleerebezem, M., J. Boekhorst, R. van Kranenburg, D. Molenaar, O. P. Kuipers, R. Leer, R. Tarchini, S. A. Peters, H. M. Sandbrink, M. W. Fiers, W. Stiekema, R. M. Lankhorst, P. A. Bron, S. M. Hoffer, M. N. Groot, R. Kerkhoven, M. de Vries, B. Ursing, W. M. de Vos, and R. J. Siezen.** 2003. Complete genome sequence of Lactobacillus plantarum WCFS1. Proc. Natl. Acad. Sci. USA **100:**1990–1995.
18. **Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer.** 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305:**567–580.
19. **Lee, S. F., and T. L. Boran.** 2003. Roles of sortase in surface expression of the major protein adhesin P1, saliva-induced aggregation and adherence, and cariogenicity of *Streptococcus mutans*. Infect. Immun. **71:**676–681.
20. **Mazmanian, S. K., G. Liu, E. R. Jensen, E. Lenoy, and O. Schneewind.** 2000. Staphylococcus aureus sortase mutants defective in the display of surface proteins and in the pathogenesis of animal infections. Proc. Natl. Acad. Sci. USA **97:**5510–5515.
21. **Mazmanian, S. K., H. Ton-That, K. Su, and O. Schneewind.** 2002. An iron-regulated sortase anchors a class of surface protein during Staphylococcus aureus pathogenesis. Proc. Natl. Acad. Sci. USA **99:**2293–2298.
22. **Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. Sigrist, R. Vaughan, and E. M. Zdobnov.** 2003. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res. **31:**315–318.
23. **Navarre, W. W., and O. Schneewind.** 1994. Proteolytic cleavage and cell wall anchoring at the LPXTG motif of surface proteins in gram-positive bacteria. Mol. Microbiol. **14:**115–121.
24. **Navarre, W. W., and O. Schneewind.** 1999. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbiol. Mol. Biol. Rev. **63:**174–229.
25. **Ni Eidhin, D., S. Perkins, P. Francois, P. Vaudaux, M. Hook, and T. J. Foster.** 1998. Clumping factor B (ClfB), a new surface-located fibrinogen-binding adhesin of Staphylococcus aureus. Mol. Microbiol. **30:**245–257.
26. **Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne.** 1997. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int. J. Neural Syst. **8:**581–599.
27. **Notredame, C., D. G. Higgins, and J. Heringa.** 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302:**205–217.
28. **Pallen, M. J., A. C. Lam, M. Antonio, and K. Dunbar.** 2001. An embarrassment of sortases—a richness of substrates? Trends Microbiol. **9:**97–102.
29. **Schneewind, O., P. Model, and V. A. Fischetti.** 1992. Sorting of protein A to the staphylococcal cell wall. Cell **70:**267–281.
30. **Sutcliffe, I. C., and R. R. Russell.** 1995. Lipoproteins of gram-positive bacteria. J. Bacteriol. **177:**1123–1128.
31. **von Heijne, G.** 1989. The structure of signal peptides from bacterial lipoproteins. Protein Eng. **2:**531–534.
32. **Womble, D. D.** 2000. GCG: the Wisconsin package of sequence analysis programs. Methods Mol. Biol. **132:**3–22.