Methodology article

LocateP: Genome-scale subcellular-location predictor for bacterial proteins

Miaomiao Zhou¹, Jos Boekhorst¹,4, Christof Francke¹,2 and Roland J Siezen*¹,2,3

Address: ¹Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands, ²TI Food and Nutrition, and Kluver Centre for Genomics of Industrial Fermentation, Wageningen, The Netherlands, ³NIZO Food Research, Ede, The Netherlands and 4Department of Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands

Email: Miaomiao Zhou - m.zhou@cmbi.ru.nl; Jos Boekhorst - j.boekhorst@uu.nl; Christof Francke - c.francke@cmbi.ru.nl; Roland J Siezen* - r.siezen@cmbi.ru.nl

* Corresponding author

Published: 27 March 2008

BMC Bioinformatics 2008, 9:173 doi:10.1186/1471-2105-9-173

This article is available from: http://www.biomedcentral.com/1471-2105/9/173

© 2008 Zhou et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: In the past decades, various protein subcellular-location (SCL) predictors have been developed. Most of these predictors, like TMHMM 2.0, SignalP 3.0, PredSi and Phobius, aim at the identification of one or a few SCLs, whereas others such as CELLO and Psortb.v.2.0 aim at a broader classification. Although these tools and pipelines can achieve a high precision in the accurate prediction of signal peptides and transmembrane helices, they have a much lower accuracy when other sequence characteristics are concerned. For instance, it proved notoriously difficult to identify the fate of proteins carrying a putative type I signal peptidase (SPIase) cleavage site, as many of those proteins are retained in the cell membrane as N-terminally anchored membrane proteins. Moreover, most of the SCL classifiers are based on the classification of the Swiss-Prot database and consequently inherited the inconsistency of that SCL classification. As accurate and detailed SCL prediction on a genome scale is highly desired by experimental researchers, we decided to construct a new SCL prediction pipeline: LocateP.

Results: LocateP combines many of the existing high-precision SCL identifiers with our own newly developed identifiers for specific SCLs. The LocateP pipeline was designed such that it mimics protein targeting and secretion processes. It distinguishes 7 different SCLs within Gram-positive bacteria: intracellular, multi-transmembrane, N-terminally membrane anchored, C-terminally membrane anchored, lipid-anchored, LPxTG-type cell-wall anchored, and secreted/released proteins. Moreover, it distinguishes pathways for Sec- or Tat-dependent secretion and alternative secretion of bacteriocin-like proteins. The pipeline was tested on data sets extracted from literature, including experimental proteomics studies. The tests showed that LocateP performs as well as, or even slightly better than other SCL predictors for some locations and outperforms current tools especially where the N-terminally anchored and the SPIase-cleaved secreted proteins are concerned. Overall, the accuracy of LocateP was always higher than 90%. LocateP was then used to predict the SCLs of all proteins encoded by completed Gram-positive bacterial genomes. The results are stored in the database LocateP-DB http://www.cmbi.ru.nl/locatep-db[1].

Conclusion: LocateP is by far the most accurate and detailed protein SCL predictor for Gram-positive bacteria currently available.
Background
In bacteria, secreted proteins are involved in stress sensing, substrate binding, cell communication, microbe-host interaction, adhesion, and other essential processes relevant to the environment and life style of the organisms. The secreted proteins are exported via various mechanisms and are retained by the bacterial cell via various interactions or released to the medium (Figure 1A). To identify the "secretome" [2] on a genome scale, subcellular proteomic studies have been carried out [3-7]. Although these experimental methods have contributed greatly to our knowledge of the subcellular location (SCL) of a variety of proteins, until now their scope has remained limited. In contrast, high-throughput computational methods for prediction of SCL sequence characteristics can be easily applied to every species whose genome has been sequenced.

Computational methods have gained considerable precision in the past decades. Initial tools focussed on detecting the presence, type and location of protein transmembrane segments, including signal peptides for targeting and translocation of proteins. One of the very first SCL prediction methods was introduced by Kyte and Doolittle [8] in 1982 with their amino acid hydropathy index. Since the late 90's machine-learning methods became more prominent, including neural networks [9-11], hidden Markov models (HMM) [12-14], support vector machines [2,15-23], Bayesian networks [24,25], and combined algorithms [26-33]. Moreover, present studies tend to combine different resources and methods [34-37]. For example, Chou et al. [38] combined gene ontology and functional domain databases, Shatkay et al. [39] combined text search and sequence data, and Marcotte et al. [40] combined protein homology and phylogenetic profiles in their studies.

Unfortunately, as a result of the trade-off between specificity and accuracy, computational methods will always be prone to error. Moreover, the number of false predictions increases even further when the SCL-related sequence characteristics have not been properly identified. For instance, among the Sec-dependent exported proteins, current predictors have severe difficulties to distinguish the proteins that are cleaved from the cell membrane by the type I signal peptidase (SPIase) – in this paper we will refer to these proteins as "secreted" – from a relatively large group of membrane-anchored proteins that also contain a putative SPIase-cleavage site but are not cleaved – in this paper we will lump these proteins in the category "N-anchored" [41-43].

As knowledge on the precise SCL of a protein is especially important to judge the biological nature and role of its activity, we constructed a new SCL prediction pipeline called LocateP. Our pipeline is geared to identify the detailed SCL of bacterial proteins by combining existing and novel prediction tools. Special effort was made to increase the accuracy of the prediction of N-anchored proteins. The version of LocateP presented here focuses on SCL prediction of proteins from Gram-positive bacteria.

Results
The construction of the SCL-prediction pipeline LocateP
A major drawback of most current sub-cellular location (SCL) predictors is that they are not aimed at the prediction of very specific SCLs but merely at the rather broad locations intracellular, membrane bound/associated and extracellular, in line with the Swiss Prot classification system. We therefore constructed a SCL predictor pipeline LocateP, that distinguishes 7 SCLs and 3 targeting pathways that can be identified in Gram-positive bacteria, with a focus on extracellular SCLs (see Figure 1A).

The LocateP pipeline was designed such that it mimics the protein secretion process in Gram-positive bacteria. The pipeline structure can be categorized as follows: (1) secretion pathway prediction, (2) transmembrane-segment detection, (3) signal peptide identification, and (4) cleavage and retention signal recognition. The LocateP pipeline employs existing SCL prediction tools (Table 1) as well as our own new and more accurate methods for the prediction of lipoproteins, Tat-secreted, N-terminally anchored, C-terminally anchored and secreted proteins (see Methods). LocateP uses at least 2 prediction methods for each SCL, in order to increase prediction accuracy. The selection criteria imposed on these methods were derived from literature. The LocateP pipeline is depicted in Figure 2; its construction is described in more detail in the "Methods" section and in the legend of Figure 2. A detailed flow chart is presented in Additional file 1.

Making the distinction between N-anchored and secreted proteins containing a SPI-cleavage site
In the past, the sequence corresponding to the signal peptide has been subdivided into three distinct regions: the N, H and C regions [28,44-46] (Figure 1B). Most of the membrane proteins with a single N-terminal TM anchor are easily identified as they do not have a predicted cleavage site for signal peptidases. However, as mentioned above, the prediction of SCL of proteins containing a putative signal peptidase type I (SPIase) cleavage site appears particularly difficult for current SCL predictors. Although many Sec-exported proteins are cleaved by the SPIase, a considerable number of proteins is not cleaved and remains membrane-anchored via the N-terminus [41].

To identify the features that determine cleavage, the multiple sequence alignments of the signal peptides from experimentally validated N-anchored and secreted pro-
Figure 1

(A): Classification of protein SCLs in Gram-positive bacteria. The secreted proteins can be divided into the following subgroups: (i) N-terminal hydrophobic tail anchored (N-anchored), (ii) C-terminal hydrophobic tail anchored (C-anchored), (iii) covalent lipid-anchored, (iv) covalently/non-covalently cell-wall anchored, (v) secreted/released (defined as proteins that are Sec/-Tat-secreted and cleaved by the signal peptidase I), and (vi) non-classically secreted/released proteins via minor pathways [120, 163]. Based on the Swiss Prot classification system the SCLs could be categorized into: Cytoplasmic, Membrane (multi-transmembrane, N-/C-anchored), Cell wall (LPxTG-anchored) and Extracellular (lipid-anchored, secreted, bacteriocin-like) proteins. (B): The structure of known signal peptides. The overall structure of Tat- and Sec-dependent signal peptides is commonly conserved as distinct consecutive N, H and C regions. The N region is the start of the protein containing positively charged residues. The H region follows the N region and is a string of consecutive hydrophobic residues which can form an α-helix in the membrane. The C region contains the signal peptidase cleavage signals. Known cleavage/retention signals include the AxAA type I SPase cleavage site [163, 172], the L-x-x-C (so-called lipobox) type II SPase cleavage site [157] and the AxA Tat-substrate cleavage site [88, 90, 173]. The LPxTG-type motif is a C-terminal sorting signal which is involved in the covalent attachment of proteins to the peptidoglycan of the cell wall. The signal peptide of proteins targeted for minor secretion pathways does not follow the N-H-C structure [2, 125, 163].
proteins [41] containing a putative SPI cleavage site in *Bacillus subtilis* were analyzed. To enhance the signal, orthologous sequences from other *Bacilli* were added in the analysis (see Materials and Methods). The Weblogo [47] of the two collections of sequences are given in Figure 3A. No distinguishing pattern could be detected by eye. Therefore, a series of HMMs were constructed based on the sequence alignments of the N-anchored and secreted proteins. Nine pairs of HMMs were built for sequences surrounding the putative SPI cleavage site. Different numbers of residues on either side of the putative cleavage site were included in the models in order to investigate the roles of the H-region and the C-region in cleavage-site recognition. When the HMM pairs were applied to the two respective sets of sequences, it appeared that the HMM pair containing an equal number of residues on either side of the putative cleavage site performed best in predicting correctly whether the cleavage site was genuine or not (Figure 3B).

**Table 1: Recent methods for protein SCL prediction**

| Speciality                                      | Tool                      | Reference |
|------------------------------------------------|---------------------------|-----------|
| Membrane protein predictor                     | a TM-HMM                  | [12]      |
| Both transmembrane helices and signal peptide predictor | a Phobius                | [14]      |
| Signal peptide predictor                       | a SignalP                 | [18]      |
|                                               | a Predisi                 | [98]      |
|                                               | signal peptidase type I cleavage site motif | [41]      |
| Lipoprotein predictor                          | b LipoP                   | [151]     |
|                                               | a signal peptidase type II cleavage site motif | [41, 157] |
| Tat-secreted protein predictor                 | b TatP                    | [86]      |
|                                               | a Tat-find.v.1.2           | [174]     |
| Protein subcellular location classifier        | b Psortb.v.2.0            | [17]      |
|                                               | b CELLO                   | [20]      |
|                                               | b Gpos-PLoc               | [28]      |
|                                               | Augur                     | [27]      |
| Minor pathway secreted protein predictor       | a Bagel                   | [149]     |
|                                               | SecretomeP 2.0            | [128]     |
| Mycobacteria protein SCL predictor             | b TBpred                  | [95]      |

a. Tools included in the LocateP pipeline
b. Tools used for comparison and validation of LocateP

**Flowchart of the LocateP pipeline.** Firstly, the possibility of being secreted by the Tat pathway was calculated by combining Tat-find v1.2 [91] and our Tat-specific HMMs (RR-HMM, CS-HMM). Bacteriocin-like proteins were identified using Bagel [149]. Secondly, Phobius [14], Predisi [98], SignalP 3.0 [18] and TM-HMM 2.0 [12] were combined to identify transmembrane regions. Those proteins without any predicted TM segments were considered intracellular, whereas those with TM segments were divided into multi-TM membrane proteins, N-anchored membrane proteins or secreted/released proteins (single N-terminal TM segment, possibly signal peptide), and C-anchored membrane proteins (signal peptide and single C-terminal TM segment). Thirdly, a sortase-substrate HMM [165] was used to distinguish LPxTG-type peptidoglycan-anchored proteins from C-anchored membrane proteins. Subsequently, signal peptidase type II (SIP) substrates were predicted by combining existing lipoprotein motif models [41, 157] and new lipoprotein HMMs. The remaining proteins were classified into the categories secreted/released or N-anchored membrane proteins. See Methods and additional file 1 for more details. Abbreviation: A-S = Anchored-Secreted; TMS = TransMembrane Segment; SP = Signal Peptide; C/N-TM = C/N-terminally transmembrane anchored; LPxTG = LPxTG cell-wall anchored.
The individual HMMs were not mutually exclusive for either of the two sets of sequences. However, when the scoring of the two HMMs of the pair was combined into a scoring matrix, the experimentally determined non-cleaved and cleaved sequences could be distinguished almost perfectly. The scoring matrix was included in LocateP; details of the matrix are described in the legend of Figure 3 and in the "Methods" section.

**Initial validation of LocateP**

Ideally, the performance of the LocateP pipeline should be checked with large experimentally validated data sets. Unfortunately, the availability of such large data sets is rather limited. LocateP was tested first with the experimental data set of Tjalsma et al. [41] which was used to create the HMM pair that distinguishes the N-anchored and secreted proteins containing a putative SPI-cleavage site. LocateP was able to distinguish these proteins with an accuracy of >90%. Then the performance of LocateP was tested on ten other data sets. These sets were extracted from literature describing other SCL prediction tools. LocateP performed extremely well on these sets, as the prediction accuracy was always higher than 90% (Table 2).

A second check was done with data collected from TransportDB [48]. Based on expert knowledge on the composition and location of various transport systems and their functional components, the SCL prediction of 1336 transport-related proteins from *Bacillus subtilis* 168, *Bacillus cereus* ATCC14579 and *Lactobacillus plantarum* WCFS1 was verified (Table 3). For a difficult group like the substrate-binding proteins of ABC transport systems, LocateP identified 113 of 124 proteins in a correct SCL for substrate binding (96 lipoproteins, 6 secreted and 11 N-anchored proteins) [49-59]. For the other groups of transport-related proteins, the predicted SCL fitted the biological role of the transport proteins in ~98% of the cases.
A third less quantitative check included a comparison of the LocateP predictions for all N-anchored and secreted proteins of the *Bacillus subtilis* genome with their NCBI functional annotations (Additional file 2, 3 and 4). Nearly all of the predictions appeared to make biological sense according to literature: most of the predicted N-anchored proteins were annotated to be involved in processes that are related to the cell-envelope, such as cell division, trans-
port, cell-envelope biogenesis, mobility, competence, signal transduction, protein turnover, etc; most predicted secreted proteins were indeed known to be secreted enzymes such as extracellular carboxydrases [60], alkaline phosphatases [61,62], metalloproteases [63], neutral proteases, and subtilisin-family proteases [64].

**Further validation and comparison of LocateP with other tools and pipelines**

Recently, Gardy et al. [44] have compared most of the current SCL classifiers, and some tools showed excellent performance. We compared the performance of LocateP to a selection of these tools, including the individual SCL predictors and other general integrative SCL classifiers that were considered best (Table 2).

*On N-anchored, secreted, lipid-anchored and multi-transmembrane protein prediction*

LocateP and several individual SCL prediction tools were applied to the same collection of reference data sets. LocateP showed similar or higher recall to PrediSi, Phobius and SignalP 3.0 at signal peptide detection, respectively (Table 2, test sets 1, 2, 3, 4). LocateP performed clearly better than all other tools at predicting lipoproteins and multi-transmembrane proteins (Table 2, test sets 6, 7, 8). For the group of N-anchored proteins, LocateP clearly outperformed all other tools with a much higher prediction specificity and accuracy (Table 2, test set 5).

As has been noted by others [12,14,42], the N-anchored membrane proteins form an ambiguous group with respect to the location of their biological activity, i.e. outside or inside the cell. Various N-anchored proteins are actually active at the cytoplasmic side of the bacterial cell membrane [65-75]. Due to the lack of reliable distinguishing algorithms and experimental data, no reliable prediction methods for these "outside-in" proteins are available yet [12,29,43,76-78]. As a result, in the current version of LocateP, proteins are only annotated as "N-terminally anchored"; most are presumed to function outside the cell, while some might have intracellular activity. A few of the known intracellular cases are indicated in Additional file 2.

*On Tat-secreted protein prediction*

Recent research pointed out that the Tat-export pathway plays an important role as a parallel protein secretion pathway to the Sec-pathway in some Gram-positive organisms [79-85]. Unfortunately, Sec-signal peptide detectors have a high false-negative prediction rate on Tat-substrates [86]. Therefore, we considered it necessary to include a Tat-secreted protein prediction tool in the LocateP pipeline, and we combined two newly created Tat-secreted protein-specific HMMs (see Methods) with Tat-find v.1.2 for the SCL prediction of these proteins. Tat-signal peptides are known to have an almost invariable double Arg or Lys+Arg motif (RR-motif) [87-90] upstream of the transmembrane segment. It appeared important that the Tat-secreted protein predictors can discriminate the Tat-signal peptides from sequences (especially transmembrane helices) that contain consecutive positively charged residues.

We compared the performance of LocateP, TatP and Tat-find v1.2 on the proteins containing a RR/RK pattern in their N-terminus (test sets 3, 4 and 9). LocateP clearly performed better than the other two specific tools when tested with intracellular and membrane proteins sets, and thus showed an excellent capability of Tat-signal peptide detection (Table 2). Moreover, it appeared that TatP and Tat-find v1.2 predicted several proteins to be secreted via the Tat-pathway in 22 species that apparently lack the relevant pathway genes [91], whereas LocateP did not find any Tat-pathway substrates in those species. Thus, LocateP showed the best overall performance among the Tat-pathway prediction tools for gram positive bacteria.

*Comparing LocateP and other integrative SCL classifiers*

According to the comparative study of Gardy et al. [44], CELLO [20] is one of the best SCL classification pipelines. We therefore evaluated the performance of LocateP as an integrative SCL classifier by comparing it to CELLO and the widely used pipeline Psortb.v.2.0 [25]. Other pipelines like SubLoc [92], LOCtree [93], Proteome Analyst [94], P-CLASSIFIER [33] and PSLpred [36] were not selected because they either do not provide prediction of membrane proteins, or are tailored for Gram-negative bacteria, or in the best case showed similar performance to Psortb.v.2.0 or CELLO. Recently, a SCL prediction tool called Gpos-Ploc [28] was published that classifies Gram-positive proteins. LocateP was not compared to Gpos-Ploc because its web server accepts only one sequence per search. Moreover, the overall accuracy of the tool is reported to be only ~85% [28].

---

**Table 3: Validation of LocateP predictions of transporter systems using the annotation in TransportDB**

| Species                  | Number of transport-related proteins with identified SCL | LocateP accuracy |
|--------------------------|--------------------------------------------------------|------------------|
| *Bacillus subtilis* 168  | 426                                                    | 98.2%            |
| *Bacillus cereus* ATCC14579 | 571                                                  | 97.5%            |
| *Lactobacillus plantarum* WCFS1 | 373                                                  | 98.8%            |

---

(http://www.biomedcentral.com/1471-2105/9/173)
LocateP had an accuracy lower than CELLO (Table 2) when tested with data extracted from the Swiss-Prot database (test set 10 [28]). However, when compared using experimental data (test sets 3, 4, 5, 6, 7), CELLO and Psortb v2.0 showed dramatically poor prediction rates (Table 2). This poor performance relates to the fact that the training data of CELLO and Psortb v2.0 were from the Swiss-Prot database (i.e., part of test set 10 and 11). This database does not distinguish between N-anchored, secreted and lipoproteins, and at the same time the members of these groups are distributed over two general classes: "membrane" and "extracellular". Thus, in essence the poor performance of CELLO and Psortv.b.2.0 is a consequence of the less-specific classification in Swiss-Prot (Table 2). Vice versa, the lower accuracy of LocateP on the Swiss-Prot data is related to the inconsistency in the classification.

TBpred [95] is a SCL classifier that was especially designed for mycobacteria, based on the idea that organism-specific methods might have higher accuracy [96,97]. We compared LocateP with TBpred using the training data of TBpred (test set 11). Surprisingly, LocateP showed considerably higher accuracy than TBpred, especially on lipoprotein and secreted protein prediction, even though no mycobacterial proteins were involved in the lipoprotein prediction training process of LocateP.

Finally, the performance of LocateP was compared to Augur [27], a computational pipeline that also combines many existing tools. Augur detects signal peptides and transmembrane helices using only SignalP and TMHMM, and consequently the accuracy of N-anchored protein prediction of Augur is much lower than with LocateP. Augur also falsely predicted 8 lipoproteins out of a test-set of 114 non-lipoproteins (test sets 4 and 5), which implied a higher false-positive rate than LocateP on lipoprotein prediction.

**Comparative analysis of protein subcellular location in Gram-positive bacteria**

LocateP was applied to the encoded proteins of all complete Gram-positive bacterial genomes available in the NCBI database. The average distribution of proteins grouped by predicted SCL was calculated for each genome. Despite the different genome sizes, Gram-positive bacteria tend to have a similar distribution of proteins over certain SCLs independent of class or family, and this independency also holds for individual Gram-positive bacterial genomes (Table 4). We note that the fractions of intracellular and membrane proteins predicted by LocateP in Gram-positive genomes were consistent with what was previously estimated by other tools [12,18,24,25]. The complete genome predictions can be viewed in our database LocateP-DB [1].

**Discussion**

Although the early SCL-prediction tools performed rather poorly, current tools perform rather well on specific categories of signal-peptide containing proteins and membrane proteins [44], reaching an accuracy of 96%. Nevertheless, for other groups like secreted, N-anchored and lipoproteins these tools still perform rather poorly. As the latter groups represent a considerable part of the secretome, we decided to design a new SCL-identification pipeline called LocateP.

The performance of LocateP was checked against the best current tools and it outperformed all of them, particularly when difficult groups of proteins and SCLs were concerned. The outstanding performance was achieved though the generation of specific HMMs based on protein sequences whose cellular fate had been experimentally tested. For instance, it has long been a problem to identify secreted and N-anchored proteins from the group of proteins carrying a putative SPI-cleavage site motif. Formerly, the H-region together with the cleavage site were considered to be the key elements of SPIase-substrate recognition. Therefore, previous signal-peptide predictors were constructed focusing on the H-region and/or on the cleavage site [14,18,26,98-100]. However, Carlos et al. [101] found that the H-region of the SPIase substrate was not critical for peptidase-cleavage capability but that, in contrast, mutations in the C-region of originally non-cleaved proteins caused alternative cleavage. They therefore claimed that specific substrate-enzyme interactions around the C-region should be decisive for SPIase-cleavage site recognition. Indeed, our analysis of the signal sequences of a group of secreted and N-anchored proteins indicated that the C-region is important, but that at the same time also the H-region carries properties that determine the protein's fate (i.e. to be or not to be cleaved). The fact that the performance of the dedicated HMMs became worse when the sequence was extended beyond 30 residues implies that the decisive information is present in this stretch of sequence. LocateP improved the separation of N-anchored and secreted proteins from >40% (by Phobius [41]) to > 90% without disturbing the SCL prediction of the other types of proteins.

LocateP was designed as a pipeline, and hence could have performed less well on specific categories than specialized tools. In particular, the performance would have been considerably lower if the flow scheme had been chosen wrongly. However, a comparison of the performance on lipoproteins, membrane proteins, Sec-secreted and Tat-secreted proteins with the specialized tools LipoP 1.0, TMHMM 2.0, Phobius, SignalP 3.0, Predisi, Tat-find v.1.2 and TatP shows that LocateP does not suffer from being a pipeline tool. Apparently, our choice to mimic the order in the bacterial secretion process was a correct one. In fact,
it has been shown by others that the SCL prediction can be improved considerably by simulating the protein sorting processes [93,102].

Overall, LocateP performed very well, with an accuracy higher than 95% for nearly all categories, and only slightly lower in one case (91% for N-anchored proteins), but still considerably better than all other tools. LocateP could be used to distinguish 7 SCLs and 3 sorting pathways and avoided the inconsistent SCL classification which most SCL classifiers inevitably inherited from Swiss-Prot.

Because of the high prediction accuracy of LocateP on proteins of known biological function (see e.g. Additional file 2), we expect that the SCL prediction of proteins of unknown function should also be equally reliable. In principle, the genome-scale SCL predictions made by

| Table 4: LocateP-predicted average distribution (%/(STDEV)) of proteins over different SCLs for Gram-positive bacteria |
|---------------------------------------------------------------|
| **Class/order level**                                         |
| Species            | Actinobacteria | Bacillales | Clostridia | Lactobacillales | Mollicutes |
| Average genome size | 4098           | 3573       | 2969       | 2048            | 724        |
| **Grouped according to LocateP classification**               |
| N-anchored (Membrane) | 5.0/(1.1)  | 5.7/(0.6)  | 6.8/(1.0)  | 5.8/(0.7)       | 8.7/(3.1)  |
| C-anchored (Membrane) | 0.3/(0.2)  | 0.1/(0.1)  | 0.2/(0.1)  | 0.2/(0.1)       | 0.3/(0.3)  |
| Multi-transmembrane (Membrane) | 16.5/(2.6) | 20.3/(1.4) | 16.9/(2.8) | 17.9/(2.1)      | 17.1/(2.3) |
| Intracellular (Cytoplasmic) | 74.3/(2.8) | 69.8/(2.2) | 73.2/(3.6) | 72.9/(2.0)      | 71.4/(3.8) |
| Lipid anchored (Extracellular) | 2.2/(0.5)  | 2.3/(0.4)  | 1.6/(0.6)  | 1.6/(0.5)       | 1.9/(1.6)  |
| Secreted (Extracellular) | 3.0/(0.9)  | 2.1/(0.5)  | 2.1/(0.5)  | 1.8/(0.6)       | 2.3/(1.3)  |
| Secreted via minor pathways (Extracellular) | 0.1/(0.1)  | 0.1/(0.1)  | 0.1/(0.1)  | 0.28/(0.2)      | 0.04/(0.1) |
| LPxTG Cell-wall anchored (Cell wall) | 0.1/(0.2)  | 0.4/(0.4)  | 0.1/(0.2)  | 0.6/(0.4)       | 0.03/(0.1) |

| **Grouped according to Swiss-Prot classification**            |
| Membrane | 21.4/(2.7) | 26.2/(1.7) | 23.8/(3.4) | 23.8/(1.9) | 26.1/(3.9) |
| Cytoplasmic | 74.3/(2.8) | 69.8/(2.2) | 73.2/(3.6) | 72.9/(2.0) | 71.4/(3.7) |
| Extracellular | 5.4/(1.1)  | 4.5/(0.7)  | 3.8/(0.8)  | 3.7/(0.8)  | 4.2/(1.9)  |
| Cell wall | 0.1/(0.2)  | 0.4/(0.4)  | 0.1/(0.2)  | 0.6/(0.4)  | 0.03/(0.1) |

| Species level |
| Organism | Spn | Lla | Sau | Lmo | Lpl | Cac | Bsu | STDEV |
| Total proteins | 2105 | 2321 | 2656 | 2846 | 3009 | 3672 | 4105 |

| **Grouped according to LocateP classification (%)** |
| N-anchored (Membrane) | 4.5 | 5.9 | 6.0 | 4.9 | 5.2 | 6.9 | 6.2 | 0.8 |
| C-anchored (Membrane) | 0.1 | 0.1 | 0.1 | 0.4 | 0.2 | 0.2 | 0.1 | 0.1 |
| Multi-transmembrane (Membrane) | 17.9 | 18.4 | 19.5 | 19.1 | 20.5 | 18.1 | 20.7 | 1.1 |
| Intracellular (Cytoplasmic) | 74.7 | 72.8 | 70.5 | 71.1 | 70.2 | 71.3 | 69.1 | 1.9 |
| Lipid anchored (Extracellular) | 1.7 | 1.4 | 2.2 | 2.0 | 1.6 | 1.7 | 2.0 | 0.3 |
| Secreted (Extracellular) | 1.2 | 1.9 | 2.1 | 1.7 | 1.9 | 2.3 | 2.6 | 0.4 |
| Secreted via minor pathways (Extracellular) | 0.5 | 0.0 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.2 |
| LPxTG cell-wall anchored (Cell wall) | 0.5 | 0.5 | 0.5 | 1.5 | 1.1 | 0.1 | 0.1 | 0.5 |

| **Grouped according to Swiss-Prot classification (%)** |
| Membrane | 22.4 | 24.4 | 25.5 | 24.4 | 25.9 | 25.2 | 27.0 | 1.4 |
| Cytoplasmic | 74.7 | 72.8 | 70.5 | 71.1 | 70.2 | 71.3 | 69.1 | 1.9 |
| Extracellular | 3.4 | 3.3 | 4.4 | 4.0 | 3.8 | 4.1 | 4.8 | 0.5 |
| Cell wall | 0.5 | 0.5 | 0.5 | 1.5 | 1.1 | 0.1 | 0.1 | 0.5 |

Abbreviations: Spn: S. pneumoniae; Lla: L. lactis; Sau: S. aureus; Lmo: L. monocytogenes; Lpl: L. plantarum; Cac: C. acetobutylicum; Bsu: Bacillus subtilis

(page number not for citation purposes)
LocateP provide an excellent starting point for functional annotation and experimental analysis of encoded proteins of unknown function, as they provide numerous clues about where to look for a certain biological activity.

Although LocateP already performs quite well, there is inevitably room for improvement. For instance, in the Swiss-Prot database, many of the annotated cell-wall proteins are secreted proteins bound to the cell surface via non-covalent interactions. Known elements of non-covalent binding include choline-binding domains, LysM domains, type 2 cell-wall binding domains, GW-modules, Lysin-binding motifs, ChW-binding motifs, Wxl domains, LPP-region binding, S-layer proteins, and others [103-117].

The current version of LocateP was designed to predict only the covalent cell-wall (peptidoglycan) binding mechanism of proteins by dedicated sortases. For instance, among the 14 cell-wall proteins in test set 10d [28], 13 are non-covalently cell-wall bound secreted proteins. LocateP correctly predicted 10 of these as "secreted proteins", but does not allow for the fact that these proteins could be cell-wall bound via non-covalent mechanisms after secretion (Table 2). Future versions of LocateP will include non-covalent binding to the cell wall.

Not all mechanisms of protein secretion or modification are known to date and not all have been included in the LocateP pipeline yet [118-123]. This is the case for proteins that have been shown to occur at various locations or those that are secreted via minor pathways. Examples are proteins that are either cleaved multi-domain proteins [124], auto-transporters found in both cytoplasmic and extracellular locations [125], or proteins with various SCL depending on growth phase and/or specific environment [124,126]. The multi-compartment proteins and minor-pathway secreted proteins appear to be rare in most bacteria, and their sorting mechanisms are not completely understood yet. Therefore, the current version of LocateP predicts only one SCL for such proteins, which may be only partially correct. In contrast, Psortb v2.0 and CELLO were claimed to be capable of multi-location prediction [126,127]. Both tools employ machine-learning methods and the predicted multiple locations should represent a certain statistical significance even without large-scale experimental evidence. However, both tools inevitably generate a considerable number of false positives. Similarly, SecretomeP 2.0 [128], which was made to predict non-classically secreted proteins, was not included in LocateP because of its high false-prediction rate. The recent predictors Euk-mPloc [129] and Hum-mPloc [97] incorporated up-to-date Eukaryotic proteins that were found to have multiple compartments and the tools achieved rather satisfying accuracies. Similar tools will be included or constructed for LocateP when more experimental data on multiple locations of bacterial proteins are available.

Another group of proteins that is not treated separately by LocateP is the group that is exported by unknown mechanisms and is known as the Gram-positive periplasmic proteins [130-133]. Carlsson et al. [134] recently reported that in Gram-positive bacteria the secreted proteins could be directed to different extracellular regions including a periplasmic space. In fact, the prediction of a subcellular location "periplasmic" in Gram-positive was not included in any published SCL prediction tools for Gram-positive bacteria, except in Gpos-PLoc [28]. However, the Gpos-PLoc prediction algorithm was based on only 5 proteins which were extracted from the Swiss-Prot database. Indeed, among these 5 proteins, four were expressed in the E. coli periplasmic space, but no evidence exists that they are also expressed in the periplasm of a Gram-positive organism [135-138]. Moreover, one protein (P29166) was proven to be located inside the cell [139]. LocateP predicted correctly that 4 of the 5 proteins should be secreted and are located outside of the plasma membrane, while P29166 was predicted to be cytoplasmic, in line with the experimental evidence.

LocateP was first tailored for the SCL prediction of Gram-positive bacterial proteins; therefore prediction of the Gram-negative specific proteins, such as β-barrel membrane proteins, was not yet included in the pipeline. LocateP was compared to SigTree [140], a signal-peptide detector based on sets of experimentally verified E. coli proteins, using the same data set from E. coli (data not shown). LocateP showed slightly lower accuracy than SigTree did, which suggests that the sequence composition of signal peptides from Gram-positive and Gram-negative bacterial proteins could be different. Future versions of LocateP will be improved and extended to Gram-negative bacterial protein SCL prediction by incorporating Gram-negative specific subcellular-location prediction tools.

Finally, we must emphasize that in several cases an automatic SCL prediction of a protein will inevitably give an incorrect prediction using LocateP or any other tool: (i) when the start codon of a gene encoding a protein with a signal peptide has been wrongly identified (either too far upstream or too far downstream), (ii) when a frame shift in the open-reading frame leads to different fragments of encoded proteins, and (iii) when an intracellular protein contains a signal peptide-like hydrophobic helix near the N-terminus; in this case such helices generally fold into the interior of the globular protein [141-144].
Conclusion
As detailed and accurate genome-scale SCL prediction of encoded proteins is highly desired by scientists in various biological research areas, numerous existing and newly developed tools were combined into one pipeline: LocateP. To date, LocateP is the most detailed protein SCL predictor for Gram-positive bacterial proteins among all tools that have been reported, in that it presently distinguishes 7 different SCLs and 3 sorting pathways, with focus on extracellular SCLs. Moreover, it is also the most accurate SCL predictor, especially on distinguishing N-anchored and secreted proteins. LocateP was applied on all completed Gram-positive bacterial genomes from the NCBI sequence database. The results are updated synchronously with GenBank updates and are publicly available via the database LocateP-DB [1]. The present version contains SCL predictions for 436,771 proteins in 148 genomes of Gram-positive bacteria. These genome-scale SCL predictions provide an excellent starting point for experimentalists to improve the functional annotation of proteins.

Methods
Sources of sequence information and location data
The genome sequences of Gram-positive bacteria were extracted from GenBank on May 30th, 2007 (*ORGANISM* annotation fields: 'Firmicutes' or 'Actinobacteria'), and were continuously updated since then. Protein sequences of Bacillus species were collected both from GenBank on April 1st, 2007 and from the ERGO database [145] on November 15th, 2006.

Eight different protein data sets of known subcellular location were selected from literature describing other tools and describing proteome studies (Table 2 legend). In order to check the performance of the LocateP pipeline, the SCL predictions were checked against an expert evaluation of the functional location of transport-related proteins from several Gram-positive bacterial genomes in TransportDB [48] on May 30th, 2007, and an expert evaluation against the protein function annotation as retrieved from GenBank on August, 10th, 2007.

Sequence analysis and evaluation of performance
Multiple sequence alignments were built with MUSCLE [146]. HMMs were built with HMMER [147]. Wherever appropriate, HMMs of varying length and different regions of the aligned N-terminal sequences of proteins were made, and the HMM that performed best was selected. Performance was evaluated using the statistical measure recall (or sensitivity) which is the number of true positives divided by the sum of the true positives and the false negatives.

Bioinformatics tools included in the LocateP pipeline
Many studies have compared and evaluated currently available transmembrane segment and signal peptide predictors [41,44,76,98-100,148]. Based on those studies and our own preliminary trials the following tools were selected to be included in our SCL prediction pipeline LocateP: TMHMM 2.0 [12], Phobius [14], SignalP 3.0 [18], PrediSi [98], and Bagel [149] (Table 1). Of these, TMHMM 2.0 and SignalP 3.0 are the most popular ones in the field; Phobius was selected for its high specificity on transmembrane segment identification; PrediSi was selected because it was trained with comparatively recent experimental data, and because it slightly outperformed SignalP 3.0 when applied to Gram-positive bacterial proteins [98]. We also included the predictor Bagel for non-classically secreted bacteriocin-like proteins [149]. The membrane protein predictor MemType-2L [150] includes topology prediction of N-anchored proteins but showed rather low accuracy with our experimental datasets; therefore this tool was not included in LocateP. Some other tools were not incorporated either because of a high false-prediction rate (e.g. HMMTOP [13] and SecretomeP 2.0 [128]), a low specificity for Gram-positive bacteria (e.g. Lipop 1.0 [151]), or simply the lack of stand-alone installable software packages (e.g. TatP [86], Signal-3L[152], Signal-CF[153] and Tat-pred [154]).

Signal peptide detection
LocateP detects signal peptides by scanning the protein N-terminus, which was defined as the initial 60 amino acids of the protein, using SignalP 3.0, Phobius and PrediSi. Some proteins have a signal peptide shortly after these 60 amino acids. These proteins were predicted as “intracellular”, but we added the extra remark of “TMH start AFTER 60” to the annotation indicating that these proteins could be secreted. No attempt was made to choose alternative start codons of incorrectly predicted start sites of ORFs.

Specific HMMs to determine the SCL of proteins with a putative SPI-cleavage site
Recently, Tjalsma et al. have experimentally determined the SCL of a large number of Bacillus subtilis proteins [41]. The experimental set contained 66 proteins with a putative SPI-cleavage site. Of these 36 appeared to be cleaved and thus secreted, whereas 30 were shown to remain N-anchored. We named these sets “EXP-secreted” and “EXP-anchored”, respectively, and used them to construct set-specific HMMs. To enhance the inherent signal, both sets were expanded by adding orthologous sequences from other Bacilli. First, homologs were searched with BLASTP [155] in the ERGO genome database [145] using full-length sequences. Only the three best BLAST hits were considered orthologs, when they also showed conserved gene context and functional annotation, high similarity and similar protein length. In this way, after removing
orthologs containing identical N-terminal sequences, 27 secreted and 23 N-anchored orthologs could be added to the "EXP-" sets.

Pairs of HMMs were built to separate the group of proteins with a putative SPI cleavage-site into those that are cleaved (i.e. secreted/released) and those that are not (i.e. N-anchored). The sequences were aligned around the putative cleavage site and the length of the HMMs was varied (length ≥8). All HMM pairs were applied to both "EXP-" sets; the E-value was set at 10,000 to assure each protein gained an HMM score. For each pair the separation between truly cleaved and truly N-anchored proteins was analyzed and it appeared that the HMM pair containing equivalent amounts of H and C region residues achieved the highest specificity in distinguishing the two sets (see Figure 3A). The most specific HMM pair had a length of 25 amino acids and ran from residue -14 to +10 relative to the cleavage site Alanine (see Figure 3A).

The individual HMMs of the selected pair (HMM_{non-cleaved}/HMM_{cleaved}) each displayed a relatively high specificity, but this was increased significantly by combining the two HMMs. A generic scoring scheme was derived via the following procedure: i) The HMM scores were rounded to discrete integers and the score distribution for the EXP-anchored and EXP-secreted protein sets was used to determine a first cut-off. The discrete HMM scores related to the HMM_{non-cleaved} ranged from -19 to +20 with all non-cleaved (i.e. N-anchored) proteins scoring higher then 3, those related to the HMM_{cleaved} ranged from -29 to +20 with all cleaved (i.e. secreted) proteins scoring higher then 0. In fact, for both HMMs only in a small scoring range the two protein groups overlapped. Therefore, all sequences with a score ≤2 using the HMM_{non-cleaved} were attributed the SCL: SEC-secreted, and those with a score ≤1 using the HMM_{cleaved} were attributed the SCL: N-anchored. ii) For those sequences that scored ≥2 with the HMM_{non-cleaved} and >1 with the HMM_{cleaved} the score with both models was compared. In case HMM_{cleaved} > HMM_{non-cleaved} score, the sequence was considered SEC-secreted, whereas, in case HMM_{non-cleaved} ≥ HMM_{cleaved} score, the sequence was considered N-anchored.

The creation and selection of a specific HMM for lipoprotein prediction

The experimental data of Tjalsma et al. [41] indicated that at least 42 distinct proteins of *Bacillus subtilis* are lipoproteins. This set of proteins was taken and expanded with orthologs from 18 closely related *Bacillus* species using an Inparanoid [156] search for best bi-directional hits. After removing the sequences which contain identical initial 50 residues, 219 putative orthologous lipoproteins could be added. As all lipoproteins are anchored to the cell membrane by thioether linkage of the conserved lipobox cysteine to a diglyceride [41,56,151,157,158], the sequences were aligned around the lipobox. Eight HMMs were built based on different N-terminal regions from these proteins varying in length between 5 and 30 residues. Each HMM was applied to the original dataset of Tjalsma et al. and the performance was evaluated. The HMM with a length of 21 residues (-20 residues to the lipobox Cysteine) showed the highest specificity when the T-score was set to 3. Gaps were allowed in this model except in the region of the lipobox (residue -5 to the lipobox Cysteine).

**The creation and selection of a specific HMM for Tat-secreted protein prediction**

The 105 putative Tat-secreted proteins (according to Swiss-Prot) from the TatP-positive training set [86] were taken as the initial set for generating Tat-specific HMMs. The sequences were aligned either around the double Arg or Lys+Arg motif (RR-motif) [87-90] upstream of the transmembrane helix or the putative AxA triplet cleavage site [86] predicted by TatP downstream of the transmembrane helix. Eleven HMMs with different lengths were generated. A combination of two HMMs was found to be most specific with the training data, together with the restriction of an E-value smaller than 10: one HMM contained 2 residues in front and 16 after the twin-arginine motif, and the other HMM contained 17 amino acids in front and 1 residue after the triplet cleavage site. Interestingly, these two HMMs partly overlapped each other by the transmembrane (H) region. According to Taylor et al. [154], the -3 to +7 residues surrounding the twin-arginine motif, and the other HMM contained 17 amino acids in front and 1 residue after the triplet cleavage site. Therefore, these two HMMs partly overlapped each other by the transmembrane (H) region. According to Taylor et al. [154], the -3 to +7 residues surrounding the twin-arginine motif, and the other HMM contained 17 amino acids in front and 1 residue after the triplet cleavage site. However, the region of the lipobox (residue -5 to the lipobox Cysteine)
order to avoid potential errors, LocateP also scans all proteins assuming them to be Sec-secreted, except for the bacteriocin-like secreted proteins. If the Tat-secreted possibility score of a protein was significant, the final subcellular location of this protein was marked "Possibly Tat-" as an extra reference.

**Specific criteria for LPxTG-anchored and C-anchored protein prediction**

The following topological criteria were used to identify LPxTG-type cell-wall anchored and C-anchored membrane proteins. For the selection of LPxTG-anchored proteins, the criteria were [165]: (i) the protein has only one N-terminal signal peptide/TM segment and only one C-terminal TM segment, (ii) the protein contains an LPxTG-type motif; (iii) the LPxTG-type motif is followed by the C-terminal transmembrane helix and a positively charged C-terminal tail. These criteria were validated with 85 experimentally verified LPxTG-anchored proteins [166-171] and 83 of them were correctly identified.

The criteria used for predicting C-anchored proteins were: (i) the protein has only 2 predicted TM helices, one situated at the N-terminus and one at the C-terminus, (ii) the protein has a cleaved N-terminal signal peptide, (iii) the protein has a C-terminal transmembrane helix and a positively charged C-terminal tail but no LPxTG motif, (iv) the distance between the N-terminus and C-terminus is larger than 45 residues.

**Authors’ contributions**

RS conceived the project and acquired the funding, MM performed the research and constructed the LocateP pipeline and database, JB helped in tool selection, CF and RS supervised the project, while MM, CF and RS wrote the manuscript. All authors read and approved the manuscript.

**Additional material**

**Additional file 1**

Flow chart and decision tree of the LocateP pipeline. The different SCL tools used at each decision step are indicated. The different SCLs distinguished are boxed in the middle; "Bacteriocin" signifies bacteriocin-like proteins secreted by non-classical pathways, identified by Bagel. a) all tools agreed, b) all possible hits, c) majority vote, d) 2–3 TM segments and C-terminus detected by LPxTG HMM.

Click here for file [http://www.biomedcentral.com/content/supplementary/1471-2105-9-173-S1.pdf]

**Additional file 2**

The LocateP predicted N-anchored and secreted proteins with known function in Bacillus subtilis.

Click here for file [http://www.biomedcentral.com/content/supplementary/1471-2105-9-173-S2.doc]

**Additional file 3**

The LocateP predicted N-anchored and secreted proteins of unknown function in Bacillus subtilis.

Click here for file [http://www.biomedcentral.com/content/supplementary/1471-2105-9-173-S3.xls]

**Additional file 4**

Literature references for other experimental evidence (Yes (O)) listed in additional file 2.

Click here for file [http://www.biomedcentral.com/content/supplementary/1471-2105-9-173-S4.doc]

**Acknowledgements**

We thank the authors of the tools used in the LocateP pipeline, especially Anne de Jong, Anders Krogh, Karsten Hiller, and Andre Billion for Bagel, Lipopor 1.0, PrediSi and Augur, respectively. We thank Antoinette Killian for stimulating discussion and Qinghu Ren for his help with TransportDB. Special thanks to the BioRange programme members Jack Leunissen for PASScan software support, Tim Hulsen for the help in database construction, Elena Marchiori and Jaap Herings for suggestions on machine-learning methods. This work is part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). Christof Frangopoulos is supported by a grant from the Kluiver Centre for Genomics of Industrial Fermentation.

**References**

1. LocateP-DB [http://www.cmbi.ru.nl/locatep-db/]
2. Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM: Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiol Mol Biol Rev* 2000, 64(3):515-547.
3. Huang F, Parmryd I, Nilsson F, Persson AL, Pakrasi HB, Andersson B, Norling B: Proteomics of Synechocystis sp. strain PCC 6803: identification of plasma membrane proteins. *Mol Cell Proteomics* 2002, 1(12):956-966.
4. Molloy MP, Phadke ND, Maddock JR, Andrews PC: Two-dimensional electrophoresis and peptide mass fingerprinting of bacterial outer membrane proteins. *Electrophoresis* 2001, 22(9):1686-1696.
5. Molloy MP, Herbert BR, Slade MB, Rabilloud T, Nouwens AS, Williams KL, Gooley AA: Proteomic analysis of the Escherichia coli outer membrane. *Eur J Biochem* 2000, 267(10):2871-2881.
6. Murakami Y, Imai M, Nakamura H, Yoshimura F: Separation of the outer membrane and identification of major outer membrane proteins from Porphyromonas gingivalis. *Eur J Oral Sci* 2002, 110(2):157-162.
7. Bumann D, Akes S, Wendland M, Janek K, Zimny-Arndt U, Sabarth N, Meyer TF, Jungblut PR: Proteome analysis of secreted proteins of the gastric pathogen Helicobacter pylori. * Infect Immun* 2002, 70(7):3396-3403.
8. Kyte J, Doolittle RF: A simple method for displaying the hydrophobic character of a protein. *J Mol Biol* 1982, 157(1):105-132.
9. Rost B, Fariselli P, Casadio R: Topology prediction for helical transmembrane proteins at 86% accuracy.  *Protein Sci* 1996, 5(8):1704-1710.

10. Sidhu A, Yang ZR: Prediction of signal peptides using bio-basis function neural networks and decision trees.  *Appl Bioinformatics* 2006, 5(1):13-19.

11. Boden M, Hawkins J: Prediction of subcellular localization using sequence-based recurrent networks.  *Bioinformatics* 2005, 21(10):2279-2286.

12. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.  *J Mol Biol* 2001, 305(3):567-580.

13. Tusnady GE, Simon I: The HMMTOP transmembrane topology prediction server.  *Bioinformatics* 2001, 17(9):849-850.

14. Kall L, Krogh A, Sonnhammer EL: A phylogenetic method for predicting subcellular localization of proteins in Gram-positive bacteria using amino acid composition.  *Bioinformatics* 2001, 17(4):477-484.

15. Chou KC, Elrod DW: Prediction of membrane protein types and subcellular locations.  *Proteins* 1999, 34(1):137-153.

16. Hua S, Sun Z: Support vector machine approach for protein subcellular localization prediction.  *Bioinformatics* 2001, 17(8):721-728.

17. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS: PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria.  *Nucleic Acids Res* 2003, 31(13):3613-3617.

18. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: Improved prediction of signal peptides: SignalP 3.0.  *J Mol Biol* 2004, 340(4):783-795.

19. Emanuelsson O: Predicting protein subcellular localization from amino acid sequence information.  *Brief Bioinform* 2002, 3(4):361-376.

20. Yu CS, Lin CJ, Hwang JK: Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions.  *Protein Sci* 2004, 13(5):1158-1165.

21. Su CY, Lee YF, Chou KHS, Su YS, Hsu WL: Protein subcellular localization prediction based on compartment-specific biological features.  *Comput Syst Bioinformatics Conf* 2006:325-330.

22. Lei Z, Dai Y: An SVM-based system for predicting protein subcellular localizations.  *BMC Bioinformatics* 2005, 6:291.

23. Taylor PD, Attwood TK: Flower power: Toward bacterial protein sub-cellular location prediction: single-class discriminant models for all gram- and gram+ compartments.  *Bioinformation* 2006, 1(8):276-280.

24. Gardy JL, Laird MR, Chen F, Roy S, Walsh CJ, Ester M, Brinkman FS: PSORTBv2: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis.  *Bioinformatics* 2005, 21(5):617-623.

25. Hawkins J, Boden M: Detecting and sorting targeting peptides with neural networks and support vector machines.  *J Bioinf Comput Bio* 2006, 4(1):1-18.

26. Billon A, Ghai R, Chakraborty T, Hain T: Augur - a computational pipeline for whole genome microbial surface protein topology prediction and classification.  *Bioinformatics* 2006.

27. Shen HB, Chou KC: Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins.  *Protein Eng Des Sel* 2007.

28. Zhou H, Zhou Y: Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method.  *Protein Sci* 2003, 12(7):1547-1555.

29. Lorena AC, de Carvalho AC: Protein cellular localization prediction with Support Vector Machines and Decision Trees.  *Comput Biol Med* 2007, 37(2):15-25.

30. Lu Z, Szafron D, Greiner R, Lu P, Wishart DS, Poulin B, Arvick J, Macdonell C, Estner R: Predicting subcellular localization of proteins using machine-learned classifiers.  *Bioinformatics* 2004, 20(4):547-556.
uptake in Staphylococcus aureus. J Bacteriol 2001, 183(17):4994-5000.

56. Screen, R. and Vitek, K: Iron-hydroxamate uptake systems in Bacillus subtilis: identification of a lipoprotein as part of a binding protein-dependent transport system. Mol Microbiol 1993, 8(1):111-121.

57. Gomez, A, Ramon, D, Sanz, P: The Bacillus subtilis lipoprotein LplA causes cell lysis when expressed in Escherichia coli. Microbiology 1994, 140 (Pt 8):1839-1845.

58. Kim MS, Shin J, Lee W, Lee HS, Oh BH: Crystal structures of RbsD leading to the identification of cytoplasmic sugar-binding proteins with a novel folding architecture. J Biol Chem 2003, 278(28):26580-26588.

59. Stenzl R, Zagarec M: Ribose utilization in Lactobacillus sakei: analysis of the regulation of the rbs operon and putative involvement of a new transporter. J Mol Microbiol Biotechnol 1999, 1(1):165-173.

60. Mole SE, Le Coq D, Aymerich S, Gonzy-Treboul G, Gay P: Characterization of the twin-arginine translocation pathway from alkalophilic Bacillus sp. KSM-K16. Appl Biochem Biotechnol 1995, 43(3):473-481.

61. Shirai T, Suzuki A, Yamane T, Ashida T, Kobayashi T, Hitomi J, Ito S: High-resolution crystal structure of M-protease: phylogenetic analysis of the high-alkaline adaptation mechanism. Protein Eng 1997, 10(6):627-634.

62. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, Borsiss R, Boursier L, Brans A, Braun M, Brignell SC, Bron S, Broillet S, Bruschi CV, Caldwell B, Casano V, Carter NM, Choi SK, Codani JJ, Connerton IF, Danchin A et al.: The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature 1997, 390(6657):249-256.

63. Wu XC, Lee W, Tran L, Wong SL: Engineering a Bacillus subtilis expression-secretion system with a strain deficient in six extracellular proteases. J Bacteriol 1991, 173(4):4952-4958.

64. Helmann JD: Deciphering a complex genetic regulatory network: the Bacillus subtilis sigmaW protein and intrinsic resistance to antimicrobial compounds. Sci Prog 2006, 89(Pt 3):243-266.

65. Helmann JD: The bacterial twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor. Proc Natl Acad Sci U S A 2006, 103(47):17927-17932.

66. Coque JJ, Liras P, Martin JF: The Bacillus subtilis division site in Bacillus subtilis: insights into the assembly of the division site. J Biol Chem 1996, 271(26):15611-15616.

67. Lee PA, Tullman-Ercek D, Georgiou G: The bacterial twin-arginine translocation pathway: a review. Mol Microbiol 2005, 55(2):394-412.

68. Caldelari I, Mann S, Crooks C, Palmer T: The Tat pathway of the plant pathogen Pseudomonas syringae is required for optimal virulence. Mol Plant Microbe Interact 2006, 19(2):200-212.

69. Berks BC: A consensus motif for the twin-arginine target signal in Tat-dependent protein secretion. Mol Microbiol 1999, 33(4):1070-1079.

70. Helmann JD: The bacterial twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor. Proc Natl Acad Sci U S A 2006, 103(47):17927-17932.

71. Lee PA, Tullman-Ercek D, Georgiou G: The bacterial twin-arginine translocation pathway. Annu Rev Microbiol 2006, 60:373-395.

72. Jongbloed JD, Antelmann H, Hecker M, Nijland R, Bron S, Airaksinen U, Pries F, Quax WJ, van Dijl JM, Braun PG: In vitro reconstitution of the division site in Bacillus subtilis. J Bacteriol 1996, 178(2):559-566.

73. Van Mellaert L, Lammertyn E, Anne AR, Pavelka MS Jr., Braunstein M: The twin-arginine translocation pathway of Mycobacterium smegmatis is functional and required for the export of mycobacterial beta-lactamases. J Bacteriol 2005, 187(22):7667-7679.

74. Caviglia F, Schipani F, Ermolaeva M, Maier MM, St Geme J, Deboy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ehrlich SD, Haas ST, Liou KL, Lu Z, Yu Z: The complete genome sequence of the Gram-positive bacterium Bacillus subtilis ATCC 6051. J Bacteriol 2004, 186(4):1322-1340.

75. Berks BC: Multiple interactions among the competence proteins of Bacillus subtilis. Mol Microbiol 2007, 62(5):454-464.

76. Yoshimura M, Asai K, Sadaie Y, Yoshikawa H: Interaction of Bacillus subtilis extracytoplasmic function (ECF) sigma factors with the N-terminus of their potential anti-sigma factors. Microbiology 2004, 150(Pt 3):591-599.

77. Helmann JD: The extracytoplasmic function (ECF) sigma factors. Adv Microb Physiol 2002, 46:47-110.

78. Scherle R, Dietrich D, Drenkard E, Dean JF: Genes for a beta-lactamase, a penicillin-binding protein and a transmembrane protein are clustered with the cephemycin biosynthetic genes in Nocardia lactamdurans. Embry J 1993, 12(2):631-639.

79. Fleischmann RD, Adams MD, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolarov JF, Nelson WC, Nymaan LA, Ermolaeva M, Salzberg SL, Delcher A, Luntzback T, Weidman J, Khouri H, Gill M, Mikula A, Bishai W, Jacobs Jr WR Jr., Venter JC, Fraser CM: Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. J Bacteriol 2002, 184(19):5479-5490.
96. Chou KC, Shen HB: Hum-mPLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 2006, 347(1):150-157.

97. Chen HC, Chou KC: Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 2007, 355(4):1006-1011.

98. Komorowski J, Groote M, Munch R, Jahn D: PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 2004, 32(Web Server issue):W375-9.

99. Zhang Z, Henzel WJ: Signal peptide prediction based on analysis of experimentally verified cleavage sites. Protein Sci 2004, 13(10):2813-2820.

100. Meine KM, Hermjakob H, Apweiler R: A comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics 2000, 16(8):741-742.

101. Carlos JL, Paetzel M, Brubaker G, Karla A, Ashwell CM, Lively MO, Holsappel S, Kiewiet R, Terpstra P, Wambutt R, Wedler T, Delepierre M, Mock M, Fouet A: Conserved anchoring mechanisms between crystal-related cell surface S-layer proteins in Gram-positive bacteria? J Biol Chem 2000, 275(49):38813-38822.

102. Taylor PD, Atwood TK, Flower DR: Combining algorithms to predict bacterial protein sub-cellular location: Parallel versus concurrent implementations. Bioinformatics 2006, 1(8):285-289.

103. Joris B, Englebert S, Chu CP, Kariyama R, Daneo-Moore L, Shockman GD, Ghynies JM: Modular design of the Enterococcus hirae muramidase-2 and Streptococcus faecalis autolysin. FEMS Microbiol Lett 1989, 50:235-237.

104. Salyers AA, Affolter K, Aparicio V, Ausubel FM: The alternative vegetative domain of type I signal peptides in substrate cleavage site selection. J Biol Chem 2000, 275(49):38813-38822.

105. Furst P, Mosch H-J, Sulz M: A protein of unusual composition from Enterococcus faecium. Nucleic Acids Res 1989, 17(16):6724.

106. Auber-Pivert E, Davies J: Biosynthesis of butirosin in Bacillus circulans NRR B312: identification by sequence analysis and insertional mutagenesis of the butB gene involved in antibiotic production. Gene 1994, 147(1):1-11.

107. Margot P, Pagni M, Karamata D: Bacillus subtilis 168 gene lytF encodes a gamma-D-glutamyl-glycine dipeptidase responsible for the lytic effect of the alternative vegetative sigma factor sigmaM. Microbiology 1999, 145 (Pt 1):57-65.

108. Noback MA, Holsappel S, Kiewiet R, Terpstra P, Wambutt R, Wedler H, Venema G, Broo S: The 172 kb prkA-addAB region from 83393 Bacillus anthracis: cloning, expression, sequence analysis and insertional mutagenesis of prkA and addAB. Gene 1994, 145 (Pt 1):57-65.

109. Steinrucker R, Biehn H, Fiedler F, Gounon P, Cossart P: Interaction between the protein INIB of Listeria monocytogenes and lipoteichoic acid: a novel mechanism of protein association at the surface of gram-positive bacteria. Mol Microbiol 1999, 34(5):902-914.

110. Cabanes D, Dehoux P, Dussurget O, Franguel L, Cossart P: Surface proteins and the pathogenic potential of Listeria monocytogenes. Trends Microbiol 2002, 10(5):238-245.

111. Sara M: Conserved anchoring mechanisms between crystalline cell surface S-layer proteins and secondary cell wall polymers in Gram-positive bacteria? Trends Microbiol 2001, 9(2):47-9: discussion 49-50.

112. Chauvaux S, Matuschek M, Beguin P: Distinct affinity of binding sites for S-layer homologous domains in Clostridium thermocellum and Bacillus anthracis cell envelopes. J Bacteriol 1999, 181(8):2455-2458.

113. Foster SJ: Cloning, expression, sequence analysis and biochemical characterization of an autotidyl of amyloid of Bacillus subtilis 168. J Biol Chem 1991, 266(51):19877-19888.

114. Nolting J, Breton G, Omelchenko MV, Makarova KS, Zeng G, Gibson R, Lee HM, Dubois J, Qiu D, Hitti J, Wolf YI, Tatusov RL, Sabathe F, Doucette-Stamm L, Soucaille P, Daly MJ, Bennett GN, Koonin EV, Smith DR: Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum. J Bacteriol 2001, 183(16):4823-4838.

115. Desvaux M, Khan A, Beaton SA, Scott-Tucker A, Henderson IR: Protein secretion systems in Fusobacterium nucleatum: genomic identification of Type IV pilination and complete Type IV pathways brings new insight into mechanisms of pathogenesis. Biochim Biophys Acta 2005, 1713(2):92-112.

116. Desvaux M, Hebraud M, Henderson IR, Pallen MJ: Type III secretion: what’s in a name? Trends Microbiol 2006, 14(4):157-160.

117. Desvaux M, Hebraud M: The protein secretion systems in Listeria: insights into bacterial virulence. FEMS Microbiol Rev 2006, 30(5):774-805.

118. Desvaux M, Dumas E, Chafsey I, Hebraud M: Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure. FEMS Microbiol Lett 2006, 258(1):1-15.

119. Desvaux M, Scott-Tucker A, Turner SM, Cooper LM, Huber D, Nataro JP, Henderson IR: A conserved extended signal peptide region directs posttranslational protein translocation via a novel mechanism. Microbiology 2007, 153(Pt 1):59-70.

120. Ghosh J, Caparon MO, Booth H: polysaccharide pyruvylase. NAD(+) glycohydrolyase in cytosolin-mediated translocation. Mol Microbiol 2006, 62(4):1203-1214.

121. Henderson IR, Cappello R, Nataro JP: Autotransporter proteins, evolution and redefining protein secretion. Trends Microbiol 2006, 14(6):255-262.

122. Henderson IR, Navarro-Garcia F, Desvaux M, Fernandez RC, Ala’Aldeen D: Type V protein secretion pathway: the autotransporter story. Microbiol Mol Biol Rev 2004, 68(4):692-744.

123. Rey S, Gardy JL, Brinkman FS: Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. BMC Genomics 2005, 6:162.

124. Feng G, Chen YC, Lu CH, Hwang JK: Prediction of protein subcellular localization. Proteins 2006, 64(3):643-651.

125. Bendtsen JD, Kiemer L, Faasblou A, Brunka S: Non-classical protein secretion in bacteria. BMC Microbiol 2005, 5:58.

126. Chou KC, Shen HB: Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular localization prediction by incorporating multiple sites. J Proteome Res 2007, 6(5):1728-1734.

127. Zuber B, Haenni M, Ribeiro T, Minning K, Lopes F, Moreillon P, Dubochet J: Granular layer in the periplasmic space of gram-positive bacteria and fine structures of Enterococcus gallinarum and Streptococcus gordonii septa revealed by cryo-electron microscopy of vitreous sections. J Bacteriol 2006, 188(18):6652-6660.

128. Mataas VR, Beveridge TJ: Cryo-electron microscopy reveals native polymeric cell wall structure in Bacillus subtilis 168 and the existence of a periplasmic space. Mol Microbiol 2005, 56(1):240-251.

129. Mataas VR, Beveridge TJ: Native cell wall organization shown by cryo-electron microscopy confirms the existence of a periplasmic space in Staphylococcus aureus. J Bacteriol 2006, 188(3):1011-1021.

130. Pooley HM, Merchante R, Karamata D: Overall protein content and induced enzyme components of the periplasm of Bacillus subtilis. Microbiol Drug Resist 1996, 2(1):9-15.

131. Carlson S, Thallhammer-Carlhelm M, Flardh K, Sandin C, Carlhelm E, Lindahl G: Signal sequence directs localized secretion of bacterial surface proteins. Nature 2006, 442(7105):943-946.

132. Rothe B, Rothe B, Rügendorf F, Schauer R: The sialidase gene from Clostridium septicum: cloning, sequencing, expression in Escherichia coli and identification of conserved sequences in sialidases and other proteins. Mol Genet 1991, 226(1-2):190-197.

133. Bahl H, Burchhardt G, Spreinat A, Haeckel K, Wienecke A, Schmidt B, Antranikian G: alpha-Amylase of Clostridium thermosol
furogenes EM1: nucleotide sequence of the gene, processing of the enzyme, and comparison of other alpha-amylases.   
Agricol. Environ. Microbiol. 1991, 57(5):1554-1559

137. Shen SH, Chretien P, Bastien L, Sil Antony SN: Primary sequence of the glucanase gene from Oerskovia xanthanolytica. Expression and purification of the enzyme from Escherichia coli.   
J Biol Chem 1991, 266(2):1058-1063

138. Rothe B, Rojek P, Frank R, Blocker H, Schauer R: Cloning, sequencing and expression of a salidase gene from Clostridium sordellii G12.   
J Gen Microbiol 1989, 135(1):3087-3096

139. Adams MW: The structure and mechanism of iron-hydrogenases. Biochim Biophys Acta 1990, 1020(2):115-145

140. Jou M, Dominique Bellin and Bastien Chevallier: Understanding the Signal Sequences with Machine Learning. In EVO 2007 proceedings Volume LNCS 4447. Springer; 2007:pp. 57-67.

141. Legler PM, Cai M, Peterkofsky A, Clore GM: Three-dimensional solution structure of the cytoplasmic B domain of the mannan receptor Mannan-binding lectin.   
J Biol Chem 2004, 279(37):3915-3921.

142. Pollastrl G, Martin AJ, Mooney C, Vullo A: Accurate prediction of transmembrane second structure and solvent accessibility by consensus combiners of sequence and structure information.   
Bioinformatics 2007, 23(1):201.

143. Gastin CL, Pieper R, Huang ST, Gebegeorgis E, Parmar PP, Clark DJ, Alami H, Papazis L, Fleischmann RD, Gill SR, Peterson SN: Proteomic profiling of cell envelope-associated proteins from Staphylococcus aureus. Proteomics 2006, 6(5):1530-1549.

144. Sun SC, Pal S, Chowdhury S, DasGupta C: 23S rRNA assisted folding of cytoplasmic maltose dehydrogenase is distinctly different from its self-folding. Nucleic Acids Res 2002, 30(11):2390-2397.

145. Overberg R, Larsen N, Walunas T, D’Souza M, Gough P, Sodom M, D’Souza J, D’Souza M: Cell envelope-associated proteins to the cell wall.   
J Biol Chem 2006, 281(52):37242-37252.

146. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput.   
Nucleic Acids Res 2004, 32(5):1792-1797.

147. Eddy SR: Profile hidden Markov models. Bioinformatics 1998, 14(9):755-763

148. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: Locating proteins in the cell using TargetP, SignalP and related tools.   
Bioinformatics 2000, 16(9):969-978.

149. deLisa MP, Tullman D, Georgiou G: Filling quality control in the export of proteins by the bacterial twin-arginine translocation (Tat) pathway. Trends Microbiol 2003, 10(10):615-620.

150. Palmer T, Sargent F, Berk BS: Export of complex cofactor-containing proteins by the bacterial Tat pathway. Trends Microbiol 2005, 13(4):173-180.

151. Robinson C, Bolhuis A: Tat-dependent protein targeting in prokaryotes and chloroplasts. Biochim Biophys Acta 2004, 1694(1-2):135-147.

152. Berk BS, Palmer T, Sargent F: Protein targeting by the bacterial twin-arginine translocation (Tat) pathway. Curr Opin Microbiol 2005, 8(2):174-181.

153. Tsalsma H, Antelmann H, Jongbloed JD, Braun PG, Darmon E, Dorenbos R, Dubois JY, Westers H, Zilen G, Quax WJ, Kuipers OP, Bron S, Hecker M, van Dijl JM: Proteomics of protein secretion by Bacillus subtilis: separating the “secrets” of the secretome.   
Microb Mol Biol Rev 2004, 68(2):207-233.

154. Jean-Luc Falcone RK Dominique Belin and Bastien Chevallier: Targeting proteins to the cell wall envelope.   
J Mol Biol 2003, 17(4):4928-4934.

155. Marraffini LA, Dedent AC, Schneewind O: Sortases and the art of anchoring proteins to the envelopes of gram-positive bacteria. Microb Mol Biol Rev 2006, 70(1):192-221.

156. Marraffini LA, Schneewind O: Targeting proteins to the cell wall envelope of sporulating Bacillus subtilis. Microb Mol Biol Rev 2006, 70(1):192-221.

157. Gaspar AH, Marraffini LA, Glass EM, Debord KL, Ton-That H, Schneewind O: Bacillus anthracis sortase A (SrtA) anchors LPXTG motif-containing surface proteins to the cell wall envelope. J Bacteriol 2005, 187(13):4646-4655.

158. Mazmanian SK, Doniach A, Schneewind O: Sortase-catalysed anchoring of surface proteins to the cell wall of Staphylococcus aureus. Mol Microbiol 2001, 40(5):1049-1057.

159. Navarre WW, Schneewind O: Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbial Mol Biol Rev 1999, 63(1):174-229.

160. Calvo P, Pucciarrelli MG, Bierne H, Cossart P, Albar JP, Garcia-Del Portillo F: Analysis of the Listeria cell wall proteome by two-dimensional nanoliquid chromatography coupled to mass spectrometry. Proteomics 2005, 5(2):432-443.

161. van Dijl JM, de Jong A, Veena Vermaa J, Venema G, Bron S: Signal peptide l of Bacillus subtilis: patterns of conserved amino acids in prokaryotic and eukaryotic type I signal peptides.   
Eur J Biochem 1999, 266(9):2819-2828.

162. Berk BS, Sargent F, Schneewind O: Targeting proteins to the cell wall envelope.   
J Mol Biol 2006, 352(2):260-274.

163. Rose RW, Bruser T, Kissingler JC, Pohlschroder M: Adaptation of protein secretion to extremely high-salt conditions by extension of the use of the twin-arginine translocation pathway.   
Proteomics 2005, 70(1):539-550.

164. Neyts A, Cabac M, Gardy JL, Laird MR, deFays K, Lambert C, Brinkman FSL: PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res 2005, 33(suppl. 1):D164-168.