Published in final edited form as:

Environ Microbiol. 2013 April ; 15(4): 983–990. doi:10.1111/1462-2920.12105.

How many signal peptides are there in bacteria?

Dmitry N. Ivankov¹, Samuel H. Payne², Michael Y. Galperin³, Stefano Bonissone⁴, Pavel A. Pevzner⁴, and Dmitrij Frishman^{1,5,*}

¹Technische Universität München, Department of Genome-Oriented Bioinformatics, 85354 Freising, Germany

²Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352, USA

³National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

⁴University of California San Diego, La Jolla, CA 92093, USA

⁵Helmholtz Zentrum Munich, National Research Center for Environment and Health, Institute for Bioinformatics, 85764 Neuherberg, Germany

Summary

Over the last five years proteogenomics (using mass spectroscopy to identify proteins predicted from genomic sequences) has emerged as a promising approach to the high-throughput identification of protein N-termini, which remains a problem in genome annotation. Comparison of the experimentally determined N-termini with those predicted by sequence analysis tools allows identification of the signal peptides and therefore conclusions on the cytoplasmic or extracytoplasmic (periplasmic or extracellular) localization of the respective proteins. We present here the results of a proteogenomic study of the signal peptides in *Escherichia coli* K-12 and compare its results with the available experimental data and predictions by such software tools as SignalP and Phobius. A single proteogenomics experiment recovered more than a third of all signal peptides that had been experimentally determined during the past three decades and confirmed at least 31 additional signal peptides, mostly in the known exported proteins, which had been previously predicted but not validated. The filtering of putative signal peptides for the peptide length and the presence of an eight-residue hydrophobic patch and a typical signal peptidase cleavage site proved sufficient to eliminate the false-positive hits. Surprisingly, the results of this proteogenomics study, as well as a re-analysis of the *E. coli* genome with the latest version of SignalP program, show that the fraction of proteins containing signal peptides is only about 10%, or half of previous estimates.

Introduction

Protein secretion is a key mechanism of the interaction between the cell and its environment. Exported proteins can be anchored in the cytoplasmic or outer membrane, retained in the periplasmic space, or released into the medium (or, in case of parasites, into the host). The functions of exported proteins include, among others, hydrolysis of extracellular polymers (proteins, nucleic acids, polysaccharides), synthesis of the cell wall and intercellular matrix, such as biofilm, modification of the host cell, and sensing of the environmental conditions. The relative proportion of secreted and cytoplasmic proteins is a useful parameter for

*Corresponding author: Dmitrij Frishman, Technische Universität München, Department of Genome-Oriented Bioinformatics, Am Forum 1, 85354 Freising, Germany. Phone: 8161-712134, fax: 8161-712186, d.frishman@wzw.tum.de.

genome-based reconstructions of the organism's behavior and metabolism, see e.g. (Galperin, 2005).

In bacteria, translocation of proteins across the cytoplasmic membrane into the periplasm or the extracellular space is typically mediated by the interaction of their N-terminal signal peptides with the protein secretion machinery. Immediately after translocation, signal peptides are proteolytically removed by leader peptidases that are located in the periplasm or attached to the outer leaf of the cytoplasmic membrane (Paetzel *et al.*, 2002). Thus, identification of signal peptides informs on whether the given protein functions within or outside the cell, and has become an important component of genome annotation.

In classical experiments, signal peptides were identified on a case-by-case basis by determining the N-terminal sequence of the mature protein using Edman degradation and then comparing it with the sequence of the respective gene. This method is laborious and time-consuming and therefore hardly appropriate for whole-genome analysis. A thorough study of *Escherichia coli* proteins by Link and colleagues (1997), which used 2D gel analysis followed by Edman degradation, mapped 12 signal peptides. These and other early identifications provided the data necessary to elucidate the common feature of signal peptides: a tripartite structure, consisting of a positively charged *n*-domain, a hydrophobic *h*-domain, and a polar *c*-domain, which contains the cleavage motif (von Heijne, 1985; 1990). These observations allowed development of effective computational methods for predicting signal peptides, such as SignalP and Phobius (Nielsen *et al.*, 1997a; 1997b; Käll *et al.*, 2004), which have become a key component of any genome analysis. Over the years, these software tools have been improved in order to increase sensitivity and performance (Nielsen *et al.*, 1999; Käll *et al.*, 2007) and to allow better recognition of non-standard signal peptides and those with non-cleavable signal sequences (Bendtsen *et al.*, 2004; 2005; Petersen *et al.*, 2011). Several other software tools for prediction of signal peptides (Fariselli *et al.*, 2003; Hiller *et al.*, 2004; Shen and Chou, 2007; Frank and Sippl, 2008; Reynolds *et al.*, 2008) and protein localization (Nakai and Horton, 1999; Gardy *et al.*, 2005; Yu *et al.*, 2010) have been developed. However, prediction methods are still usually trained on a relatively small set of experimentally confirmed targeting sequences from a narrow taxonomic range of organisms and therefore their reliability in predicting signal peptides of as yet unknown type(s) remains suspect (Payne *et al.*, 2012). There is an obvious need for new high-throughput experimental approaches for identifying signal peptides that would provide a critical mass of reliable data for further progress of the popular software tools.

During the last decade, combined fractional diagonal liquid chromatography in conjunction with mass-spectrometry has been used to define N-terminal peptides in human and two halophilic bacteria (Gevaert *et al.*, 2003; Aivaliotis *et al.*, 2007; Staes *et al.*, 2011). Proteogenomics (using mass spectroscopy to identify proteins predicted from genomic sequences) has emerged as a promising approach to genome annotation, particularly for high-throughput identification of protein N-termini, a task that is not fully solved by the existing gene-calling algorithms, see e.g. (Frishman *et al.*, 1998). In addition, a comparison of the experimentally determined N-termini with those predicted by sequence analysis tools allows identification of the signal peptides and prediction of the extracytoplasmic (periplasmic or extracellular) localization of the respective proteins (Gupta *et al.*, 2007; Venter *et al.*, 2011). Proteogenomics approach reports a peptide cleaved from the protein N-terminus by observing spectra matching to the peptides with non-tryptic N-termini (each such peptide represents a potential new N-terminus of the protein after signal peptide cleavage). However, not all such peptides are true signal peptides and some filtering is needed to remove artifacts and obtain the set of likely signal peptides. Proteogenomics studies of *Shewanella oneidensis* (Gupta *et al.*, 2007), *Yersinia pestis* (Payne *et al.*, 2010), a

microbial biofilm community (Erickson *et al.*, 2010), and the fungus *Aspergillus niger* (Braaksma *et al.*, 2010) revealed hundreds of signal peptide cleavage events.

Given the success of the proteogenomics studies in discovering signal peptides, we sought to validate this experimental technique on classical model organisms with extensive functional annotation. Such validation aimed to (i) evaluate the reliability and specificity of the identification of N-terminal peptides through proteogenomics and (ii) compare it to the reliability of the widely used computational tools. In other words, we tried to evaluate the utility of the proteogenomic approach as a means of the genome-scale identification of bacterial signal peptides, particularly for investigation of yet unexplored organisms (Payne *et al.*, 2012). In addition, we tried to address a more fundamental question: what fraction of the total *E. coli* proteome have signal peptides and are therefore destined for export outside the cytosol?

We present here an analysis of putative signal peptides for *Escherichia coli* K-12 strain MG1655 (Blattner *et al.*, 1997; Riley *et al.*, 2006). According to the UniProt database (The UniProt Consortium, 2012), *E. coli* has the highest number of experimentally confirmed signal peptides among all bacteria. Moreover, almost 600 *E. coli* proteins have experimentally determined cellular localization (Lopez-Campistrous *et al.*, 2005). We show here that a single proteogenomics experiment recovered more than one third of all experimentally known signal peptides from *E. coli*. In addition, it experimentally confirmed a number of previously predicted signal peptides. Overall, after appropriate filtering, all putative signal peptides reported by proteogenomics appeared to correspond to actual signal peptide cleavages previously reported or predicted for *E. coli* (albeit with six of them having non-standard cleavage sites). Early analyses projected that at least 15–20% of proteins in gram-negative bacteria should have signal peptides (Nielsen *et al.*, 1997a). For *E. coli*, this fraction was estimated to be in the range of 17–32% (Hiller *et al.*, 2004; Käll *et al.*, 2004; Shen and Chou, 2007). Using proteogenomics data, we argue that this was an overestimation, and that the actual number of proteins with signal peptides in *E. coli* is likely to be substantially smaller, about 10%, which is consistent with the estimates from the latest version of SignalP.

Generation and analysis of proteogenomic data

The proteomic data were generated following a previously described protocol (Payne *et al.*, 2010), see Supplementary Materials for details. Briefly, proteins purified from a whole-cell lysate of *E. coli* K-12 strain MG1655 were digested into peptides with trypsin (see Lewis *et al.*, 2010). These peptides were analyzed by MS/MS and the resulting spectra were matched to peptide sequences from the *E. coli* reference protein set in the RefSeq database (Pruitt *et al.*, 2012) using the InsPecT program (Tanner *et al.*, 2005), followed by re-scoring with PepNovo (Frank, 2009) and MSGF (Kim *et al.*, 2008). Protein identifications were remapped to UniProt. For each identified protein, the most N-terminal peptide was selected; the 488 proteins whose peptide was not produced by tryptic digestion (i.e. was not preceded in the protein sequence by a Lys or Arg residue) were analyzed further (see Fig. 1 and Table S1). Peptides missing only the N-terminal methionine were discarded. For the remaining 129 proteins, the sequence between the start and the first observed peptide was designated the putative signal peptide. These 129 peptides (Table S2) were sorted and filtered for an appropriate length (18–36 amino acids) and checked for the presence of a hydrophobic patch of at least eight consecutive residues (for filtering details, see Supplementary Materials).

As a result of this filtering, the initial list of 488 was trimmed to a set of 96 putative signal peptides (Table S3), which were compared with the data stored in the UniProt, SPdb, and EcoGene databases (Choo *et al.*, 2005; The UniProt Consortium, 2012; Zhou and Rudd,

2012) and with predictions of the SignalP program (Bendtsen *et al.*, 2004). UniProt (<http://www.uniprot.org/>) annotates signal peptides (as well as protein localization and function) and categorizes them as experimentally proven (by direct protein sequencing), inferred by similarity, or potential (e.g. predicted by both Phobius and SignalP; see <http://www.uniprot.org/manual/signal>). SPdb database (<http://proline.bic.nus.edu.sg/spdb>, last updated in 2009) also provides information on both experimentally confirmed or predicted signal peptides, whereas EcoGene (<http://ecogene.org/>) provides a detailed manually curated annotation of *E. coli* genes, including the positions of known or predicted cleavage sites. For *E. coli* K-12, the current lists of experimentally verified signal peptides in UniProt, SPdb, and EcoGene include 156, 66, and 144, peptides, respectively, for a total of 163. Accordingly, each putative signal peptide identified by proteogenomics was checked to see (i) whether a signal peptide of the same length was reported by UniProt, SPdb, or EcoGene database or predicted by SignalP, and (ii) if it exhibited a typical signal peptidase I cleavage pattern [AGILSV]x[AGS] at positions -3, -2, or -1 relative to the cleavage site (Payne *et al.*, 2010).

Efficiency and caveats of the proteogenomic approach

Comparison of the set of 96 putative signal peptides obtained in this work (Table S3) against the 156 experimentally confirmed signal peptides from UniProt showed 58 peptides that were common between these two datasets. Thirty-one more peptides, while not experimentally confirmed until now, were consistent with SignalP and Phobius predictions, as listed in UniProt (of which 24 were also consistent with UniProt annotations, see Table S3). Twenty-two of the 31 respective proteins were localized either in the periplasm or in the outer membrane, for 8 proteins the cellular location remained unknown, and just one, potassium transporter KefA, was an integral cytoplasmic membrane protein (in UniProt, membrane-anchored periplasmic proteins AmpH, DacC, NlpA, YfhM, and YgiM were listed as localized in the cell membrane). The functions of these proteins included transport, stress response, and biofilm formation. One more identified peptide had a typical cleavage site but slightly different length than that predicted by SignalP and Phobius. The six remaining peptides came from known periplasmic proteins but differed in length from those listed in UniProt and/or predicted by SignalP and had atypical cleavage sites (Table S3); these peptides still remain to be properly validated. Thus, a single proteogenomic experiment revealed more than a third (58 out of 156) of all previously validated signal peptides and affirmed several new ones. Ninety peptides out of 96 were consistent with previous experimental data and/or predictions of the SignalP program and/or had typical cleavage sites.

Despite its efficiency in identifying signal peptides, the proteogenomic approach has a number of caveats. First, by discarding tryptic peptides, the filtering procedure necessarily removes those signal peptides that end in Lys or Arg. Fortunately, the number of such experimentally validated signal peptides in bacteria appears to be miniscule. There is an experimentally studied case of cytochrome oxidase subunit 2 (QOX2_BACCR) in *Bacillus cereus* with the signal peptide apparently ending in a Lys residue (Contreras-Zentella *et al.*, 2003). Other examples of such bacterial signal peptides listed in UniProt (beta-lactamases BLA1_BACCE and BLAC_STRAL, listeriolysin TACY_LISMO, and staphylococcal enterotoxin ETXE_STAAU) do not look credible and probably reflect additional proteolysis of the source proteins prior to Edman degradation. However, there are more than 80 such cases reported in eukaryotes, e.g. yeast asparaginase ASP21_YEAST (Kim *et al.*, 1988). Some of these cases might be genuine, others may reflect sample degradation by endogenous trypsin-like proteases, which would also be relevant for future studies.

Second, the signal peptide mapping procedure could be affected by the action of cellular peptidases other than signal peptidases I and II. In the example of the periplasmic osmotically-inducible protein Y (OsmY) shown on Fig. 1, six different peptides starting from E29 supported a 28-aa signal peptide with a perfect cleavage site, which was also consistent with the earlier results of Link *et al.* (Link *et al.*, 1997). However, detection of a single peptide with the cleavage point after M16 (Fig. 1) led to the erroneous conclusion that the N-terminal signal peptide was 16 amino acids long, which was below the 18–36 aa filtering window and therefore resulted in the exclusion of OsmY from the final set (Table S2). In addition to OsmY, the analyzed set of 129 N-terminal peptides contained four more instances (DegP, DppA, HdeA, and LivJ) of apparent false-negative assignments, caused by detection of short signal peptides along with those of correct size.

Third, some short(er) N-terminal peptides could arise from tryptic hydrolysis of full-length proteins that had not yet been processed by the signal peptidases. This results in a false-negative assessment similar to what was observed for OsmY. For example, for the well-known (and abundant) outer membrane proteins OmpC and OmpF, many MS/MS identified peptides supported the canonical signal peptides starting at the positions 22 and 23. However, we also identified other upstream MS/MS peptides that could only originate from an unprocessed protein. Detection of these peptides has been registered as evidence of the absence of N-terminal processing of OmpC and OmpF, again resulting in false-negative assignments.

Finally, we have observed several instances where the signal peptide cleavage sites reported by MS/MS differed from experimentally determined and/or predicted ones (Table S3). The reasons for these discrepancies are not clear at this time and deserve further investigation.

Evolution of signal peptide prediction

Our next step was to evaluate the fraction of all signal peptides of *E. coli* detected in the proteogenomics experiment. To our great surprise, we have come across vastly different estimates of the total number of signal peptides in the *E. coli* proteome. The first proteome-wide estimate of the number of signal peptides was given in the seminal paper introducing the SignalP method (Nielsen *et al.*, 1997a). Signal peptides were identified in 330 out of 1680 (20%) of the *Haemophilus influenzae* proteins, leading the authors to estimate that ~15–20% of any Gram-negative proteome would possess signal peptides. The next release of the SignalP program, SignalP 2.0, came in two variants, which relied on hidden Markov models and on neural network and predicted 20% and 32%, respectively, of the *E. coli* proteome to possess signal peptides (Nielsen *et al.*, 1997b; 1999). However, more recent updates of the SignalP program reported substantially lower fractions of *E. coli* proteins with signal peptides, 14–17 % for the two variants of SignalP 3.0 (Bendtsen *et al.*, 2004) and even lower, ~10%, for the more recent version, SignalP 4.0 (Petersen *et al.*, 2011). We have confirmed these estimates by running various versions of SignalP against the current UniProt reference set of 4,303 *E. coli* K-12 proteins (Fig. 3). To a significant degree, this could be attributed to the better discrimination between cleavable and non-cleavable signal peptides by the latter versions of the program (Petersen *et al.*, 2011).

Other signal peptide prediction tools exhibited a similar behavior. PrediSi (Hiller *et al.*, 2004) predicted cleavable signal peptides in 1075 (~25%) *E. coli* sequences, Signal-3L (Shen and Chou, 2007) predicted 1479 *E. coli* proteins (34% of the total proteome) to be secreted. Two other programs, Phobius (Käll *et al.*, 2004) and the later Philius (Reynolds *et al.*, 2008) came up with much smaller numbers, predicting signal peptides in 17–19 % of *E. coli* proteins.

We attempted to provide an independent estimate of the number of signal peptides in *E. coli* by comparing the proteogenomics results with the lists of predicted signal peptides. As mentioned above, among those signal peptides listed as “potential” in UniProt, 24 predictions have been fully confirmed and for 14 more proteins our data indicated a different cleavage site (Table S3). For two proteins with “potential” signal peptide, the ethanolamine utilization protein EutM and uncharacterized protein YeeZ, MS/MS detected N-terminal tryptic peptides, MEALGMIETR and MKKVAIVGLGWLGMPLAMSLAR, respectively, suggesting that they do not have (cleavable) signal peptides. For EutM, indeed, experimental data, including the 3D structure of the full-length protein, confirm that this protein is not being processed and stays in the cytoplasm, albeit in a separate carboxysome-like microcompartment (Tanaka *et al.*, 2010). We could only find some tentative data for YeeZ, suggesting that it might act outside the cell (Lynch *et al.*, 2007). Thus, EutM appears to be the only false-positive prediction in the UniProt list; YeeZ could be another one or a false-negative case like OmpC and OmpF discussed above. As a result, the UniProt list of “potential” signal peptides appears to be very close to reality, and the full set of 156 confirmed and 337 predicted signal peptides, accounting for 11.5% of *E. coli* proteins, can be considered the best available estimate. This means that the proteogenomics experiment reported here detected ~20% of all *E. coli* signal peptides.

Signal peptide prediction for other organisms

The downward trend in the estimates of the number of signal peptides is also seen for other organisms, such as the Gram-positive bacterium *Bacillus subtilis* (Fig. 3) and the yeast *Saccharomyces cerevisiae* (data not shown). As already noticed by Leversen and Wiker (2012), for mycobacteria successive versions of SignalP predict fewer and fewer signal peptides, apparently gradually approaching the true number. The improvement achieved by SignalP 4.0 over previous versions of SignalP can primarily be explained by the explicit discrimination between signal peptides and signal anchors, but also by a substantially larger dataset of experimentally known signal peptides available for training (e.g., 2,935 Gram-negative instances in SignalP 4.0 vs 334 in SignalP 3.0).

It should be noted that, despite the extremely high accuracy of SignalP predictions for *E. coli*, the fraction of false-positive and false-negative predictions could be substantially higher for lesser-studied organisms, where few signal peptides have been confirmed by experiment. For example, comparison of SignalP 4.0 predictions with proteogenomics data for the bacteria *Caulobacter crescentus* and *Rhodobacter sphaeroides* (Venter *et al.*, 2011) revealed 21 discrepancies out of 429 (5%) and eight out of 354 (2%), respectively (data not shown).

In conclusion, proteogenomics proved to be an efficient high-throughput approach to signal peptide discovery, recovering in a single experiment more than one-third of all experimentally validated signal peptides and ~20% of all signal peptides in *E. coli*. The criteria for filtering putative signal peptides (absence of C-terminal Lys or Arg, typical length, and the existence of a hydrophobic patch), while being purely empirical, allowed removal of all false-positives. Future analyses will show if these criteria could be modified in order to improve sensitivity without losing the selectivity of the signal peptide recognition. We also note that the total number of signal peptides for *E. coli* proved to be on the order of 10%, much lower than previously published estimates, but consistent with the predictions of the most recent version of SignalP program (Petersen *et al.*, 2011). Similarly low (lower than assumed previously) numbers of signal peptides could be expected also in other microorganisms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Natalya Bogatyreva for assistance. DNI was supported by the DFG International Training and Research Group RECESS (Regulation and Evolution of Cellular Systems). SHP was supported by an NSF award EF-0949047. MYG was supported by the NIH Intramural Research Program at the National Library of Medicine. Data from the Pacific Northwest National Laboratory was obtained with funding from NIAID IAA-Y1-A1-8401 in the Environmental Molecular Sciences Laboratory, a U.S. Department of Energy/Biological and Environmental Research national scientific user facility. Pacific Northwest National Laboratory is operated for the DOE by Battelle under Contract DEAC05-76RLO 1830.

References

- Aivaliotis M, Gevaert K, Falb M, Tebbe A, Konstantinidis K, Bisle B, et al. Large-scale identification of N-terminal peptides in the halophilic archaea *Halobacterium salinarum* and *Natronomonas pharaonis*. *J Proteome Res.* 2007; 6:2195–2204. [PubMed: 17444671]
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.* 2004; 340:783–795. [PubMed: 15223320]
- Bendtsen JD, Kiemer L, Fausboll A, Brunak S. Non-classical protein secretion in bacteria. *BMC Microbiol.* 2005; 5:58. [PubMed: 16212653]
- Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, et al. The complete genome sequence of *Escherichia coli* K-12. *Science.* 1997; 277:1453–1462. [PubMed: 9278503]
- Braaksma M, Martens-Uzunova ES, Punt PJ, Schaap PJ. An inventory of the *Aspergillus niger* secretome by combining *in silico* predictions with shotgun proteomics data. *BMC Genomics.* 2010; 11:584. [PubMed: 20959013]
- Choo KH, Tan TW, Ranganathan S. SPdb--a signal peptide database. *BMC Bioinformatics.* 2005; 6:249. [PubMed: 16221310]
- Contreras-Zentella M, Mendoza G, Membrillo-Hernandez J, Escamilla JE. A novel double heme substitution produces a functional *bo*₃ variant of the quinol oxidase *aa*₃ of *Bacillus cereus* Purification and paratol characterization. *J Biol Chem.* 2003; 278:31473–31478. [PubMed: 12805383]
- Erickson BK, Mueller RS, VerBerkmoes NC, Shah M, Singer SW, Thelen MP, et al. Computational prediction and experimental validation of signal peptide cleavages in the extracellular proteome of a natural microbial community. *J Proteome Res.* 2010; 9:2148–2159. [PubMed: 20218729]
- Fariselli P, Finocchiaro G, Casadio R. SPElip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics.* 2003; 19:2498–2499. [PubMed: 14668245]
- Frank AM. A ranking-based scoring function for peptide-spectrum matches. *J Proteome Res.* 2009; 8:2241–2252. [PubMed: 19231891]
- Frank K, Sippl MJ. High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics.* 2008; 24:2172–2176. [PubMed: 18697773]
- Frishman D, Mironov A, Mewes HW, Gelfand M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* 1998; 26:2941–2947. [PubMed: 9611239]
- Galperin MY. A census of membrane-bound and intracellular signal transduction proteins in bacteria: Bacterial IQ, extroverts and introverts. *BMC Microbiol.* 2005; 5:35. [PubMed: 15955239]
- Gevaert K, Goethals M, Martens L, Van Damme J, Staes A, Thomas GR, Vandekerckhove J. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat Biotechnol.* 2003; 21:566–569. [PubMed: 12665801]
- Gupta N, Tanner S, Jaitly N, Adkins JN, Lipton M, Edwards R, et al. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res.* 2007; 17:1362–1377. [PubMed: 17690205]

- Hiller K, Grote A, Scheer M, Munch R, Jahn D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* 2004; 32:W375–W379. [PubMed: 15215414]
- Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol.* 2004; 338:1027–1036. [PubMed: 15111065]
- Käll L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.* 2007; 35:W429–W432. [PubMed: 17483518]
- Kim KW, Kamerud JQ, Livingston DM, Roon RJ. Asparaginase II of *Saccharomyces cerevisiae* Characterization of the ASP3 gene. *J Biol Chem.* 1988; 263:11948–11953. [PubMed: 3042786]
- Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res.* 2008; 7:3354–3363. [PubMed: 18597511]
- Leversen NA, Wiker HG. Improved signal peptide predictions in mycobacteria? *Tuberculosis (Edinb).* 2012; 92:291–292. [PubMed: 22484109]
- Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol.* 2010; 6:390. [PubMed: 20664636]
- Link AJ, Robison K, Church GM. Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis.* 1997; 18:1259–1313. [PubMed: 9298646]
- Lopez-Campistrous A, Semchuk P, Burke L, Palmer-Stone T, Brokx SJ, Broderick G, et al. Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth. *Mol Cell Proteomics.* 2005; 4:1205–1209. [PubMed: 15911532]
- Lynch SV, Dixon L, Benoit MR, Brodie EL, Keyhan M, Hu P, et al. Role of the *rapA* gene in controlling antibiotic resistance of *Escherichia coli* biofilms. *Antimicrob Agents Chemother.* 2007; 51:3650–3658. [PubMed: 17664315]
- Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 1997a; 10:1–6. [PubMed: 9051728]
- Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst.* 1997b; 8:581–599. [PubMed: 10065837]
- Nielsen, H.; Krogh, A. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6). Menlo Park, California: AAAI Press; 1998. Prediction of signal peptides and signal anchors by a hidden Markov model; p. 122-130.
- Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* 1999; 12:3–9. [PubMed: 10065704]
- Paetzel M, Karla A, Strynadka NC, Dalbey RE. Signal peptidases. *Chem Rev.* 2002; 102:4549–4580. [PubMed: 12475201]
- Payne SH, Huang ST, Pieper R. A proteogenomic update to *Yersinia*: enhancing genome annotation. *BMC Genomics.* 2010; 11:460. [PubMed: 20687929]
- Payne SH, Bonissone S, Wu S, Brown RN, Ivankov DN, Frishman D, et al. Unexpected diversity of signal peptides in prokaryotes. *mBio.* 2012; 3:e00339-12.
- Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods.* 2011; 8:785–786. [PubMed: 21959131]
- Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* 2012; 40:D130–D135. [PubMed: 22121212]
- Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput Biol.* 2008; 4:e1000213. [PubMed: 18989393]
- Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, et al. *Escherichia coli* K-12: a cooperatively developed annotation snapshot--2005. *Nucleic Acids Res.* 2006; 34:1–9. [PubMed: 16397293]
- Shen HB, Chou KC. Signal-3L: A 3-layer approach for predicting signal peptides. *Biochem Biophys Res Commun.* 2007; 363:297–303. [PubMed: 17880924]

- Staes A, Impens F, Van Damme P, Ruttens B, Goethals M, Demol H, et al. Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat Protoc.* 2011; 6:1130–1141. [PubMed: 21799483]
- Tanaka S, Sawaya MR, Yeates TO. Structure and mechanisms of a protein-based organelle in *Escherichia coli*. *Science.* 2010; 327:81–84. [PubMed: 20044574]
- Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, et al. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem.* 2005; 77:4626–4639. [PubMed: 16013882]
- The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2012; 40:D71–D75. [PubMed: 22102590]
- Venter E, Smith RD, Payne SH. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PLoS One.* 2011; 6:e27587. [PubMed: 22114679]
- von Heijne G. Signal sequences. The limits of variation. *J Mol Biol.* 1985; 184:99–105. [PubMed: 4032478]
- von Heijne G. The signal peptide. *J Membr Biol.* 1990; 115:195–201. [PubMed: 2197415]
- Zhou J, Rudd KE. EcoGene 3.0. *Nucleic Acids Res.* 2013; 41:D613–D624. [PubMed: 23197660]

Fig. 1.

Mapping of the peptides identified by mass-spectrometry to the OsmY open reading frame (OSMY_ECOLI). The most N-terminal of the identified non-tryptic peptides starts from Leu-17, which is why the signal peptide is assumed to be 16 aa long, outside the selected 18–36 aa window, resulting in a false-negative assignment. The correct start of the mature protein is Glu-28, as in six other peptides and also in (Link *et al.*, 1997), see text for details.

Fig. 2.
The pipeline for classifying putative signal peptides deduced from proteogenomics results into signal, non-signal, and questionable ones.

Fig. 3. Evolution of signal peptide prediction

Predictions of the signal peptides in the reference proteomes of *Escherichia coli* (solid columns) and *Bacillus subtilis* (striped columns) by successive versions of the SignalP program (Nielsen *et al.*, 1997a; 1997b; Nielsen and Krogh, 1998; Nielsen *et al.*, 1999; Bendtsen *et al.*, 2004; Petersen *et al.*, 2011) (left group), and by Phobius, PredSi, Signal-3L, and Philius programs (Hiller *et al.*, 2004; Käll *et al.*, 2004; Shen and Chou, 2007; Reynolds *et al.*, 2008) (right group), ordered by the year of the software release. SignalP, versions 3.0 and 4.0, (<http://www.cbs.dtu.dk/services/SignalP/versions.php>) and Phobius (<http://phobius.sbc.su.se/>) programs were downloaded from their respective web sites and run locally; SignalP 2.0, PrediSi (<http://www.predisi.de/>) and Philius (<http://www.yeastrc.org/philius/>) were used at their respective web sites; predictions by Signal-3L were downloaded from <http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/Data.htm>.