RESEARCH ARTICLE

Unexpected Diversity of Signal Peptides in Prokaryotes

Samuel H. Payne,[a] Stefano Bonissone,[b] Si Wu,[c] Roslyn N. Brown,[a] Dmitry N. Ivankov,[d] Dmitrij Frishman,[d,e] Ljiljana Paša-Tolić,[c] Richard D. Smith,[a] and Pavel A. Pevzner[b,f]

Division of Biological Sciences, Pacific Northwest National Laboratory, Richland, Washington, USA[a]; Bioinformatics Program, University of California—San Diego, La Jolla, California, USA[b]; Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington, USA[c]; Department of Genome-Oriented Bioinformatics, Technical University of Munich, Munich, Germany[d]; Helmholtz Center Munich-German Research Center for Environmental Health, Institute of Bioinformatics and Systems Biology, Neuherberg, Germany[e]; and Department of Computer Science and Engineering, University of California—San Diego, La Jolla, California, USA[f]

ABSTRACT **Signal peptides are a cornerstone mechanism for cellular protein localization, yet until now experimental determination of signal peptides has come from only a narrow taxonomic sampling. As a result, the dominant view is that Sec-cleaved signal peptides in prokaryotes are defined by a canonical AxA motif. Although other residues are permitted in the motif, alanine is by far the most common. Here we broadly examine proteomics data to reveal the signal peptide sequences for 32 bacterial and archaeal organisms from nine phyla and demonstrate that this alanine preference is not universal. Discoveries include fundamentally distinct signal peptide motifs from *Alphaproteobacteria*, *Spirochaetes*, *Thermotogae* and *Euryarchaeota*. In these novel motifs, alanine is no longer the dominant residue but has been replaced in a different way for each taxon. Surprisingly, divergent motifs correlate with a proteome-wide reduction in alanine. Computational analyses of ~1,500 genomes reveal numerous major evolutionary clades which have replaced the canonical signal peptide sequence with novel motifs.**

IMPORTANCE **This article replaces a widely held general model with a more detailed model describing phylogenetically correlated variation in motifs for Sec secretion.**

Received 7 September 2012 **Accepted** 24 October 2012 **Published** 20 November 2012

Citation Payne SH, et al. 2012. Unexpected diversity of signal peptides in prokaryotes. mBio 3(6):e00339-12. doi:10.1128/mBio.00339-12.

Editor Stephen Giovannoni, Oregon State University

Address correspondence to Samuel H. Payne, samuel.payne@pnnl.gov.

Common to all cellular life, signal peptides play a foundational role in protein sorting and localization (1). In bacteria, the Sec signal peptide is a ~20- to 30-residue N-terminal sequence that targets proteins for export through the Sec system (2). The initial description of the signal peptide was derived from a relatively small number of protein sequences spread across bacteria and *Eukaryota* (3, 4) and consists of an N-terminal basic region, a hydrophobic patch, and a three-residue motif for signal peptidase I (SPaseI) cleavage. In prokaryotes, the cleavage motif is often termed "AxA" for the dominant use of alanine (5). Although other uncharged residues are permitted in the cleavage motif, alanine is by far the most common (6).

Clade-specific differences in the signal peptide have been investigated on a limited basis. In 1989, von Heijne and Abrahmsén analyzed five bacteria and found that the signal peptides of bacteria were similar, which has become the prevailing opinion (6). Variability within an organism has been explored (reviewed in reference 7), for example, examining how a longer N-terminal region in one particular protein affects its transport and folding. However, such research does not address diversity in the general sense. Specifically, are there organisms which utilize a divergent signal peptide motif in bacteria or archaea? Proteolytic cleavage creates a new protein N terminus that is amenable to discovery via proteomics. On this premise, we analyzed 140 million tandem mass spectra from 32 organisms from nine phyla to experimentally identify signal peptide sequences and search for novel motifs. The data identified six organisms in four phyla with fundamentally distinct sequence characteristics for their signal peptides. Computational predictions of the ~1,500-organism RefSeq collection suggest numerous phylogenetic clades that harbor novel variants.

RESULTS

Mature N termini created *in vivo* from enzymatic cleavage (such as signal peptidase cleavage of the signal peptide) are amenable to discovery from tandem mass spectrometry (MS/MS) data (see Materials and Methods). Proteomics data from 30 organisms from eight phyla were analyzed with the prokaryotic proteogenomic pipeline to comprehensively identify peptides and mature protein N termini (see Table S1 in the supplemental material). The signal peptidase I (SPaseI) motif for an organism was visualized by aligning residues −3 to +2 for all proteins where proteomics identified signal peptide cleavage. Figure 1 depicts the SPaseI motifs from organisms in seven phyla: *Proteobacteria*, *Cyanobacteria*, *Chlorobi*, *Deinococcus-Thermus*, *Spirochaetes*, *Actinobacteria*, and the archaeal phylum *Euryarchaeota*. The expected AxA is easily recapitulated with the identified sequences from *Escherichia coli* (Fig. 1A) and most other organisms (see Fig. S1). However, in five organisms we observe for the first time a departure from this canonical motif. Two *Alphaproteobacteria* (*Ehrlichia chaffeensis* and *Pelagibacter ubique*, Fig. 1C and D, respectively) have a novel motif, for which alanine at −3 and −1 is

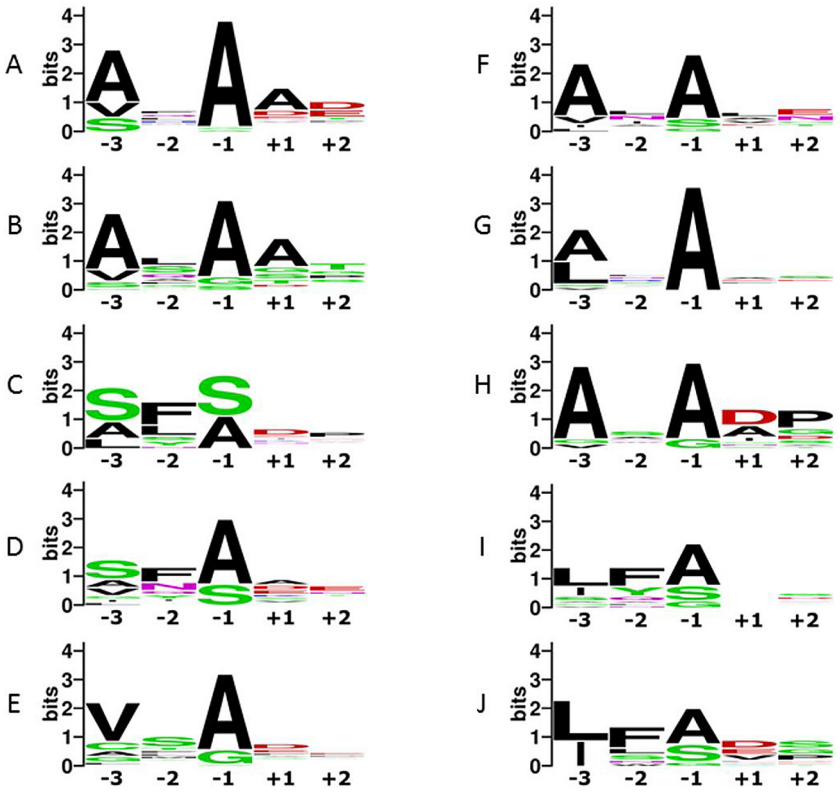

FIG 1 Different signal peptidase cleavage motifs. Signal peptide cleavage motifs are shown from 10 organisms. Residues shown are the residues −3, −2, −1, +1, +2, with cleavage being between −1 and +1. Motifs in panels C, D, E, I, and J are a significant departure from the traditional AxA signature. (A) *Escherichia coli* K-12 ($n = 85$); (B) *Deinococcus radiodurans* ($n = 56$); (C) *Ehrlichia chaffeensis* ($n = 18$); (D) *Pelagibacter ubique* ($n = 28$); (E) *Methanospirillum hungatei* ($n = 42$); (F) *Cyanothece* strain ATCC 51142 ($n = 31$); (G) *Chlorobaculum tepidum* ($n = 19$); (H) *Mycobacterium tuberculosis* ($n = 17$); (I) *Borrelia burgdorferi* ($n = 25$); (J) *Leptospira interrogans* ($n = 12$). Plots generated by http://weblogo.berkeley.edu.

marginalized. Instead, serine is dominant at −3 in both organisms and is also dominant at −1 in *E. chaffeensis*. Seven other *Alphaproteobacteria* in the data set retain the traditional motif (see Fig. S1). Both organisms from the *Spirochaetes* phylum (*Leptospira interrogans* and *Borrelia burgdorferi*, Fig. 1I and J) exhibit the same pattern of deviation. At −3, alanine is not observed but is instead replaced by leucine and isoleucine. Position −1 is occupied by both alanine and serine. The only archaeon in the data set, *Methanospirillum hungatei*, has a third distinct motif (Fig. 1E). The use of valine at position −3 for *M. hungatei* is similar to eukaryotic motifs found in humans and yeast, although computational predictions have previously suggested a more bacterium-like motif (8). Finally, although the variability of positions −3 and −1 is the most obvious deviation from the AxA rule, phenylalanine at position −2 in the divergent bacterial motifs is also significant. It appears that for these organisms, −2 is not an unrestricted position.

As the SPaseI cleavage motif is only one of three components of the signal peptide, we next investigated whether perhaps other components had changed in the organisms with a novel cleavage motif. Previous characterizations of the hydrophobic patch note this region as being primarily composed of alanine and leucine (9). However, the marginalization of alanine in the cleavage motif suggested a potential change. Using the 500 validated signal peptides from nine *Alphaproteobacteria*, we compared amino acid frequencies for the hydrophobic patch by species. Alanine, which is common in the proteome, is typically overrepresented in hydrophobic patches and comprises ~30 to 35% of patch residues for the seven organisms with a normal SPaseI motif (Fig. 2). *P. ubique* and *E. chaffeensis* contain only 5% and 7% alanine, respectively, or approximately 6 standard deviations below the median. Alanine is replaced with phenylalanine and isoleucine, whose combined increase is greater than 11 standard deviations above the levels in the typical alphaproteobacteria. When the data are normalized by background amino acid distribution in the genome, these deviations are still far outside the norm (see Fig. S2 in the supplemental material). Thus, these two organisms that deviate from the classical AxA motif also have a coordinated change in the amino acid composition of their hydrophobic patch. This phenomenon—hydrophobic patches with low alanine and high phenylalanine and isoleucine—holds for all five organisms with novel signal peptidase I cleavage motifs.

While *P. ubique* and *E. chaffeensis* have exceptionally low alanine usage in hydrophobic patches, they also have low alanine in the proteome as a whole. The proteome-wide amino acid composition of the nine *Alphaproteobacteria* shows a clear disparity between organisms which possess a normal SPaseI motif and those which possess a divergent SPaseI motif. The seven *Alphaproteobacteria* with a traditional motif have ~12 to 15% alanine proteome-wide, while *P. ubique* and *E. chaffeensis* have an alanine content of ~5%. Concomitant with the decrease of alanine is a proportionally equal increase in phenylalanine and isoleucine. Expanding these results further, we analyzed the proteome-wide composition of alanine against isoleucine and phenylalanine for all *Alphaproteobacteria* with a complete genome sequence; two distinct populations arise (see Fig. S3 in the supplemental material). Most *Alphaproteobacteria* have high alanine content (11% to 15%) and cluster with the seven organisms that have traditional signal peptides. The small number of organisms clustered at lower alanine content (4% to 6%) primarily consisted of members of the order *Rickettsiales*, which includes both *P. ubique* and *E. chaffeensis*.

To see if this was a general principle—that organisms with low global alanine usage also have variant signal peptides—all complete prokaryotic genomes were searched for proteins containing signal peptide motifs derived from our proteomics data (see Materials and Methods). Surprisingly, the search revealed many organisms with a strongly decreased use of the canonical AxA, yet they did not match well to the new motifs discovered through proteomics. In some cases, entire phyla appear to utilize novel motifs, e.g., *Spirochaetes*, *Fusobacteria*, and *Aquificae*. On the other hand, there are also organisms which deviate from very close phy-

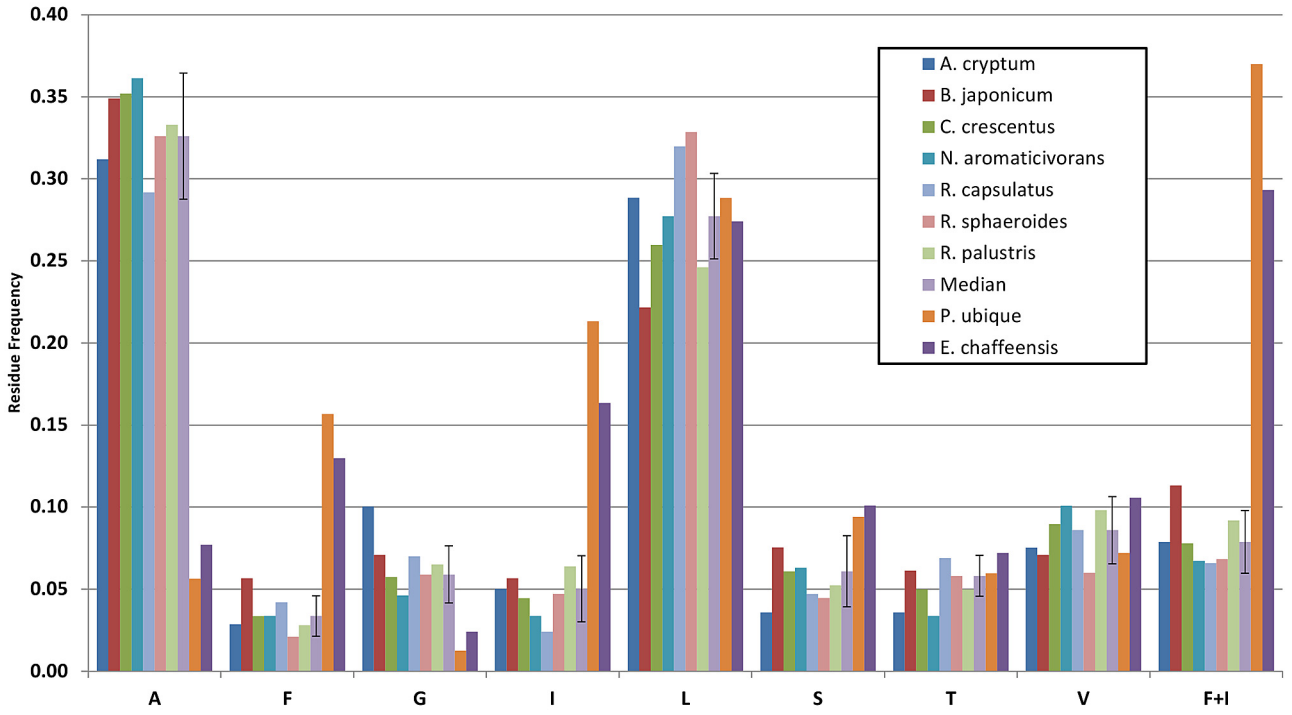

FIG 2 Amino acid frequency in the hydrophobic patch. Amino acid frequencies of the hydrophobic patch segment of signal peptides discovered through proteomics are shown for nine *Alphaproteobacteria* (*Acidiphilium cryptum*, *Bradyrhizobium japonicum*, *Caulobacter crescentus*, *N. aromaticivorans*, *Rhodobacter capsulatus*, *Rhodobacter sphaeroides*, *Rhodopseudomonas palustris*, *P. ubique*, and *E. chaffeensis*). "Median" is the median value of all nine organisms and is shown with error bars representing the standard deviation estimated from the median absolute deviation. Columns correspond to amino acids, with F+I being the sum of phenylalanine and isoleucine. Residues not commonly found in membrane-spanning regions are omitted for simplicity (values near zero in all organisms).

logenetic relatives. One example is the aphid symbiont *Buchnera*, a member of the *Enterobacteriaceae*. Several *Enterobacteriaceae* presented here and elsewhere (*E. coli*, *Salmonella enterica*, *Yersinia pestis*) all contain the AxA motif, while *Buchnera* does not.

We further investigated these hypotheses with two additional analyses. First, we validated the signal peptide cleavage motifs above, which originated from bottom-up proteomics data, with intact protein mass spectrometry (i.e., top-down). Whole proteins from periplasmic enriched samples and whole-cell lysates were analyzed, and the resulting protein identifications were processed to characterize the signal peptide cleavage (see Materials and Methods). Motifs discovered from top-down data of *Novosphingobium aromaticivorans*, *Salmonella enterica* serovar Typhimurium 14028 S, and *Shewanella oneidensis* matched what was discovered from bottom-up data. Second, we identified signal peptides from the human pathogen *Bartonella henselae*. In *Alphaproteobacteria*, where the distribution of alanine content is bimodal (see Fig. S3 in the supplemental material), *B. henselae* is one of the few organisms with medium alanine content. This can therefore show when alanine content is sufficiently low to favor novel signal peptide sequence formation. Proteomics data identified 20 proteins with signal peptide cleavage. The SPaseI motif appears to be in the process of transition (see Fig. S1). Alanine is still the plurality residue at position −3, yet it has a low information content corresponding to significant heterogeneity. Curiously, the hydrophobic patch has already deviated from the traditional alanine/leucine-rich sequence in favor of leucine/phenylalanine/isoleucine. This result suggests that the hydrophobic patch signature is the least constrained and quickest to change as alanine content in the proteome drops. This is similar to some cyanobacteria, e.g., *Prochlorococcus*, which has a variant hydrophobic patch and a changing SPaseI motif. These organisms depict the transition of the signal peptide signature from the canonical to the novel.

As a final analysis, we attempted to delineate between alanine and genomic percent GC. The translation codons for alanine are GCN, and thus, alanine composition can be roughly approximated from percent GC. To show that the driver of motif switching is global alanine levels, we chose to characterize the signal peptide sequences of the hyperthermophile *Thermotoga maritima*, which has an atypically low alanine content relative to genomic GC content: alanine, 5.8%; genomic GC, 46%. (Median alanine for organisms with 46% GC is 8.6%, and the median absolute deviation is 1.0%. Median percent GC for organisms with 5.8% alanine is 33%, and the median absolute deviation is 3%.) From proteomics data, 16 signal peptide cleavage events were identified. The observed "LFA" SPaseI motif is indeed a deviation from the traditional AxA motif (see Fig. S1 in the supplemental material).

In an attempt to identify *de novo* what the new sequences might be, we tabulated the frequency of all amino acid residue pairs i and $i + 2$ from the proteome. For a coherent view of bacterial and archaeal diversity, residue pairs were ranked by frequency in a proteome and the rankings were correlated with single amino acid frequencies across all organisms in RefSeq (see Materials and Methods). The pairs clustered into two prominent groups (Fig. 3). AxA, AxL, and LxA clustered and correlated with higher alanine content; these pairs correspond to the canonical signal peptide, both the hydrophobic patch and the SPaseI cleavage motif. Nu-

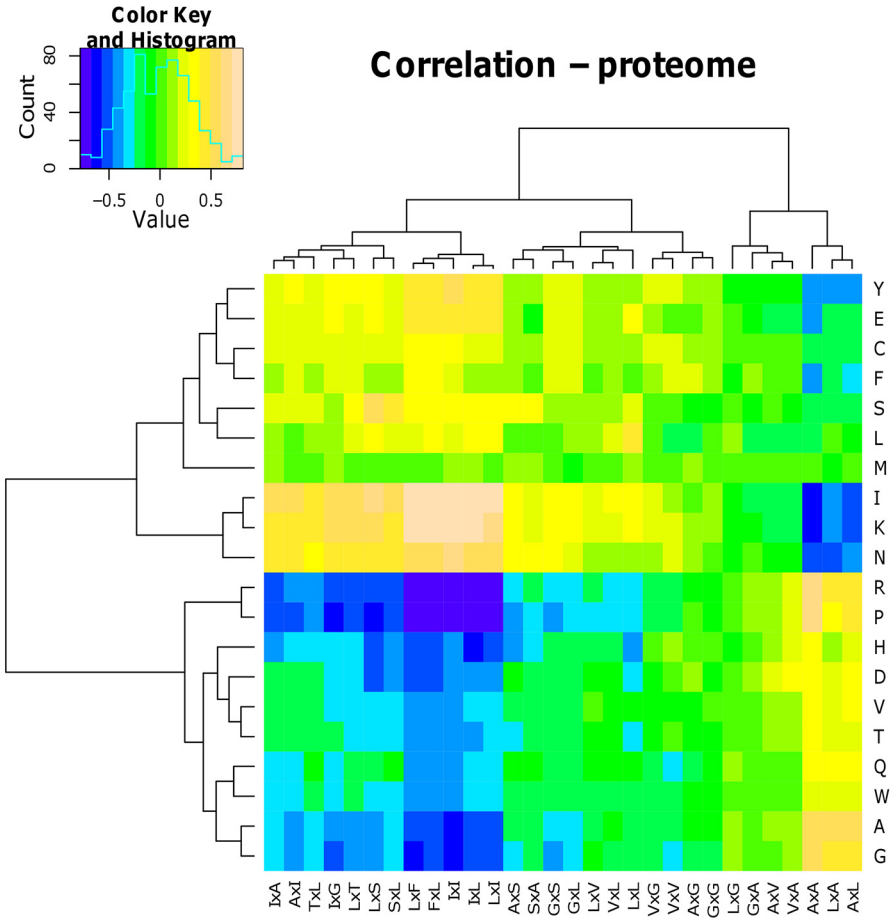

FIG 3 Diversity of residue pairs segregating by proteome composition. The mean frequency rank of diresidue pairs was correlated with the amino acid frequencies for 1,489 proteomes (see Materials and Methods). The tight cluster containing LxA, AxL, and AxA correlates with high alanine levels and describes the hydrophobic patch and SPaseI cleavage motif of canonical signal peptides. The other main branch contains pairs that correlate with low alanine levels and show the hydrophobic patch composition (e.g., LxI and FxL) or the SPaseI motif (LxS, SxL, SxA, and AxS) for novel signal peptides. In this cluster are numerous pairs which potentially contribute to yet-uncharacterized SPaseI motifs (e.g., TxL).

merous pairs clustered with low alanine. Some of these pairs are purely hydrophobic and represent hydrophobic patches, e.g., IxL and LxF. Other pairs in this second cluster contribute to novel SPaseI cleavage motifs discovered through proteomics, e.g., LxS as in the two *Spirochaetes*. Although this cluster contains the residue pairs shown here to be utilized in the novel SPaseI motifs, it also contains many pairs not yet known to contribute to the cleavage motif, e.g., TxL. These pairs potentially point to a large and uncharacterized diversity of signal peptidase I cleavage motifs.

DISCUSSION

Since its initial discovery and characterization, the signal peptide sequence in bacteria and archaea has been presumed to be universal. Through extensive and diverse proteome sampling, we show several taxa with fundamentally distinct motifs. The presence of novel motifs correlates with a global, proteome-wide reduction in alanine use. Computational predictions suggest that numerous novel motifs remain to be discovered in low-alanine organisms.

It is interesting that the novel SPaseI cleavage motifs are not similar between taxa; each taxonomic group reported here has its own consistent variant. Such clade-specific motifs highlight the need for phylogeny to be explicitly considered in sequence motifs. Sequence homology of the SPaseI, signal peptide peptidase (SPP), or signal recognition particle (SRP) genes in the organisms with a novel motif does not support horizontal gene transfer, but rather their closest homologs are from taxonomically close organisms. In the case of *E. chaffeensis* and *P. ubique*, the closest homologs are other *Rickettsiae* followed by other *Alphaproteobacteria* with normal signal peptides. We propose that the reduction of alanine in the proteome favors a deviation from the traditional signal peptide sequence and that the multiplicity of novel motifs is the result of independent evolution as opposed to horizontal gene transfer.

MATERIALS AND METHODS

Bottom-up proteomics data acquisition and analysis. Proteomics data sets for 31 organisms were created at the Pacific Northwest National Laboratory; protein purification and spectrum acquisition were performed as previously described (10, 11). Data for *Leptospira interrogans* were publicly available at the PeptideAtlas repository (PAe000349 to PAe000352 and PAe000401). Peptides were identified from tandem mass spectra using the prokaryotic proteogenomic pipeline (12). Briefly, mass spectra were searched by Inspect (13) against a translation of the genome and subsequently rescored using PepNovo (14) and MSGF (15). Searches did not include any posttranslational modifications, and in order to identify protein maturation events, we did not require tryptic specificity. We in-

TABLE 1 Parameters for signal peptide domains used for whole-genome searches

Organism	n-domain			h-domain		c-domain		
	Minimal charge	Minimal length	Maximal length	Minimal length	Minimal hydrophobicity	Minimal length	Maximal length	Minimal cleavage site probability
E. coli	1	2	13	7	10.2	4	11	2.0e−4
Borrelia	0	2	13	5	14.4	4	10	5.0e−3
Leptospira	0	3	16	7	11.1	3	10	1.4e−2
Pelagibacter	1	2	12	8	19.6	4	9	3.7e−3
Methanospirillum	0	2	14	10	15.4	3	8	2.5e−4
Ehrlichia	0	2	9	9	25.9	5	9	6.7e−3

cluded a decoy database (shuffled protein sequences) to help measure the relative peptide false-discovery rate, even though the final scoring algorithm (MSGF) does not use the decoy hits to calculate its probability values. Using the default pipeline cutoff of 1e−10 for peptide spectrum matches, we achieved a ~0.3% peptide false-discovery rate for each data set (spectrum false-discovery rate, <0.05%). Peptide-spectrum matches can be obtained from http://omics.pnl.gov/pgp/overview.php.

Proteolytic cleavage, such as signal peptide cleavage, creates a new protein N terminus that is amenable to discovery via proteomics. Although there is some variability in the residue immediately prior to cleavage (i.e., residue −1) as discussed below, the basic residues arginine and lysine are not observed in this position. Thus, the cleavage by signal peptidase can be assumed to be distinct from trypsin. Therefore, the new N terminus formed by signal peptidase cleavage produces atypical peptides which are easily identified in tryptic bottom-up proteomics. Signal peptide cleavage was identified using the prokaryotic proteogenomic pipeline as previously described (12). First, we identified all proteins whose first peptide is nontryptic on its N terminus and within 15 to 35 residues of the start methionine. This set was filtered by requiring the three expected parts of a signal peptide: early basic residue(s), a hydrophobic patch, and a three-residue recognition motif for signal peptidase I (3). Residue requirements for the cleavage motif followed those proposed by Perlman and Halvorson (3), namely, position −3 allows Ile, Val, Leu, Ala, Gly, Ser while position −1 allows Ala, Gly, Ser. Reported data sets contain at least 12 positive observations of signal peptide cleavage. Residues −3 through +2 were extracted and input to http://weblogo.berkeley.edu to visualize the cleavage motif.

Top-down data acquisition and analysis. Intact protein mass spectrometry was performed on periplasm enriched samples from *Novosphingobium aromaticivorans* and *Shewanella oneidensis*. The periplasmic fraction was released from cells by osmotic shock. Cells were washed in 50 mM Tris-Cl (pH 8), suspended in 15 ml spheroplasting buffer (50 mM Tris-Cl, pH 8, 250 mM sucrose, 2.5 mM EDTA) for 5 min, and harvested. The cell pellet was suspended in 5 mM $MgSO_4$ for 10 min, and the periplasmic fraction was collected as the supernatant after centrifugation. Centrifugal spins were performed at 10,000 × *g* for 10 min at 4°C. Periplasmic proteins were concentrated using Amicon Ultra 0.5-ml centrifugal filters with a 3-kDa-molecular-mass cutoff (Millipore, Billerica, MA), further concentrated to dryness using a Savant SpeedVac concentrator (Thermo, Fisher Scientific, Waltham, MA), and suspended in deionized water. Whole-cell protein extraction for *Salmonella enterica* serovar Typhimurium 14028 S was done as previously described but without trypsin proteolysis (16).

The intact protein reverse-phase liquid chromatography (RPLC) separation was performed on a Waters NanoAquity system with a column (80 cm by 75-μm inside diameter [i.d.]) packed in-house with Phenomenex Jupiter particles (C5 stationary phase, 5-μm particle diameter, 300-Å pore size). Mobile phase A was composed of 0.5% acetic acid, 0.01% trifluoroacetic acid (TFA), 5% isopropanol, 10% acetonitrile (ACN), and 69.75% water. Mobile phase B consisted of 0.5% acetic acid, 0.01% TFA, 9.9% water, 45% isopropanol, and 45% ACN. The operating flow rate was 0.4 μl/min. The RPLC system was equilibrated with 100% mobile phase A for 5 min and then increased to 20% mobile phase B in 1 min. A 250-min linear gradient was set from 20% mobile phase B to 55% mobile phase B. Mass spectrometry (MS) analysis was performed using an LTQ Orbitrap Velos ETD mass spectrometer (Thermo Scientific, San Jose, CA) outfitted with a custom electrospray ionization (ESI) interface. Electrospray emitters were custom made using 150-μm-outside-diameter (o.d.) by 20-μm-i.d. chemically etched fused silica (17). The heated capillary temperature and spray voltage were 275°C and 2.2 kV, respectively. For the liquid chromatography-tandem mass spectrometry (LC-MS/MS) run with ETD fragmentation, a parent spectrum was collected at a 60K resolution and was followed by high-resolution ETD MS/MS scans of the 8 most intense ions from the parent scan. Fourier transform (FT) MS/MS employed 35% normalized collision energy for CID. For the LC-MS/MS run with HCD fragmentation, a parent spectrum was collected at a 60K resolution and was followed by high-resolution ETD MS/MS scans of the 8 most intense ions from the parent scan. FT MS/MS employed 35% normalized collision energy for CID. Mass calibration was performed prior to analysis according to the method recommended by the instrument manufacturer.

Intact MS/MS data were subjected to data analysis and protein identification using MSAlign+ (18). The false-positive estimation for protein-spectrum matches was done by searching all top-down spectra against the human Uniprot database. A final cutoff of E value below e−04 was chosen for statistically significant matches. MS-Align+ produces an output of proteins with modified N termini that might be the product of signal peptide cleavage. These results were tested in a manner similar to that of the test used for bottom-up proteomics. The putative signal peptide region was required to be of an appropriate length (15 to 35 residues long) and to contain early basic residues, a hydrophobic patch, and the three-residue recognition motif for signal peptidase I cleavage.

Whole-proteome searches using proteomically derived motifs. The following methodology was employed to search for potential signal peptides within protein sequences from an organism. First, the N-terminal sequence of a protein was considered a potential signal peptide if its length was from 15 to 35 amino acid residues and it could be divided into n-, h-, and c-domains that satisfy the parameters given in Table 1. Additional requirements included the following: (i) the h-domain should not have charged residues and (ii) the n-domain should have at least one charged residue. The h-domain was considered to begin immediately after the last charged residue of the n-domain. The boundary between h- and c-domains was placed to maximize the sum of the number of hydrophobic residues in the h-domain and hydrophilic residues in the c-domain (a residue was considered hydrophobic if it had positive hydrophobicity in the Kyte-Doolittle scale [19]; otherwise, it was considered hydrophilic). When several positions give the same sum, the algorithm prefers the leftmost position, which results in shorter h- and longer c-domains. (We did not limit the process by the maximal c-domain length, which sometimes resulted in a c-domain longer than the allowed limit, and therefore, the sequence was rejected.) When several alternative cleavage sites were identified for the same protein, the one having the maximal probability was

chosen. The probability of the cleavage site was calculated from the observed MS frequency of the three residues preceding the cleavage site.

The additional parameters as well as those given in Table 1 were derived from the manual alignment of signal peptides observed in MS experiments performed separately for each organism. (We excluded from the *Methanospirillum* alignment protein YP_503815.1 having the h-domain GGQGGVTAA, deviating by hydrophobicity and length.) For all six organisms, the sequences of the signal peptides observed in MS were successfully recognized by the algorithm, except for one sequence for *Borrelia* having a longer n-domain than allowed (NP_212304.1) and five sequences for *E. coli* having c-domains longer than allowed (NP_417066.1, NP_416192.1, NP_417701.1, NP_414694.1, and NP_415941.5).

***De novo* whole-proteome searches.** To predict the two residues present at the signal peptidase I cleavage site in all organisms, we looked for frequently occurring pairs of amino acids. For every position i in the protein, the residues at i and $i + 2$ are tabulated. This results in the frequency of the 400 di-amino-acid pairs as separated by a nonconstrained residue. For example, the sequence AxA would be binned with the diresidue pair AA; TxA would be binned with TA and so forth. We tabulated only from proteins that contained a hydrophobic patch in their first 30 residues. Calculating the hydrophobic patch followed the same methodology as discussed above.

In *E. coli*, the most common pair was AA. When the frequency of pairs was plotted against the position in the protein, clear signals arose. The red line for AA peaks sharply around residue 20, where the AxA signal peptidase I cleavage motif is commonly observed. The three pairs LL, LA, and AL are all abundant over the broad range of 5 to 20, correlating perfectly with the hydrophobic patch location and composition (see Fig. S4 in the supplemental material). Using this methodology, we looked at all organisms in the RefSeq collection. As an example of an organism with a noncanonical signal peptide, *L. interrogans* is shown in Fig. S5. The pairs frequently seen at positions 5 to 20 included LL, IL, LI, II, FL, LF, and not AL or LA, a finding which reflects the distinct composition of the hydrophobic patch. The residue pair LS spiked near the motif site, where AA was conspicuously absent.

To identify novel motifs in organisms, we first rank all diresidue pairs in each of the 1,489 organisms based on the maximal count within the signal peptide range. For our purposes, the signal peptide range is defined as positions 10 to 40 of a protein. This rank was augmented in the following way: any residue pair, whose frequency did not spike above the chi-square significance test, was moved to the lowest possible rank (maxRank + 1). This was done so that residue pairs which have a distinct locational preference were ranked higher than those whose location was indiscriminate. The empirical frequencies of amino acids were also computed and ranked for each organism based on the proteome. Pearson's correlation was then computed between the rank of each diresidue pair and single amino acid rank. To aid in visualization, only the 30 diresidues with the best mean rank were plotted against the 20 amino acids, resulting in a 20-by-30 heat map plot. The columns were then clustered based on similarity.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00339-12/-/DCSupplemental.

Figure S1, TIF file, 1.3 MB.
Figure S2, TIF file, 0.3 MB.
Figure S3, TIF file, 0.2 MB.
Figure S4, TIF file, 0.1 MB.
Figure S5, TIF file, 0.1 MB.
Table S1, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Josh Adkins and Penny Colton (Pacific Northwest National Laboratory) for discussions related to the manuscript.

This work was supported by an NSF award to S.H.P. (EF-0949047). Data from the Pacific Northwest National Laboratory were obtained 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.

We declare no competing financial interests.

REFERENCES

1. **Blobel G, Dobberstein B.** 1975. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J. Cell Biol. **67**:835–851.
2. **von Heijne G.** 1990. The signal peptide. J. Membr. Biol. **115**:195–201.
3. **Perlman D, Halvorson HO.** 1983. A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J. Mol. Biol. **167**:391–409.
4. **von Heijne G.** 1983. Patterns of amino acids near signal-sequence cleavage sites. Eur. J. Biochem. **133**:17–21.
5. **Paetzel M, Karla A, Strynadka NC, Dalbey RE.** 2002. Signal peptidases. Chem. Rev. **102**:4549–4580.
6. **von Heijne G, Abrahmsén L.** 1989. Species-specific variation in signal peptide design. Implications for protein secretion in foreign hosts. FEBS Lett. **244**:439–446.
7. **Hegde RS, Bernstein HD.** 2006. The surprising complexity of signal sequences. Trends Biochem. Sci. **31**:563–571.
8. **Bardy SL, Eichler J, Jarrell KF.** 2003. Archaeal signal peptides—a comparative survey at the genome level. Protein Sci. **12**:1833–1843.
9. **Nielsen H, Engelbrecht J, Brunak S, von Heijne G.** 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10**:1–6.
10. **Angel TE, et al.** 2010. Proteome analysis of Borrelia burgdorferi response to environmental change. PLoS One **5**:e13800. http://dx.doi.org/10.1371/journal.pone.0013800.
11. **Frank AM, et al.** 2011. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. Nat. Methods **8**:587–591.
12. **Venter E, Smith RD, Payne SH.** 2011. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One **6**:e27587. http://dx.doi.org/10.1371/journal.pone.0027587.
13. **Tanner S, et al.** 2005. InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal. Chem. **77**:4626–4639.
14. **Frank AM.** 2009. A ranking-based scoring function for peptide-spectrum matches. J. Proteome Res. **8**:2241–2252.
15. **Kim S, Gupta N, Pevzner PA.** 2008. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. J. Proteome Res. **7**:3354–3363.
16. **Ansong C, et al.** 2011. Experimental annotation of post-translational features and translated coding regions in the pathogen Salmonella typhimurium. BMC Genomics **12**:433. http://dx.doi.org/10.1186/1471-2164-12-433.
17. **Kelly RT, et al.** 2006. Chemically etched open tubular and monolithic emitters for nanoelectrospray ionization mass spectrometry. Anal. Chem. **78**:7796–7801.
18. **Liu X, et al.** 2012. Protein identification using top-down. Mol. Cell. Proteomics **11**(6):.008524. http://dx.doi.org/10.1074/mcp.M111.008524.
19. **Kyte J, Doolittle RF.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.