COMPUTATIONAL PREDICTION AND COMPARATIVE ANALYSIS OF PROTEIN SUBCELLULAR LOCALIZATION IN BACTERIA

by

Jennifer Leigh Gardy
B.Sc., University of British Columbia, 2000
Graduate Certificate in Biotechnology, McGill University, 2001

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In the
Department of Molecular Biology and Biochemistry

SIMON FRASER UNIVERSITY

Spring 2006

APPROVAL

Name: **Jennifer Leigh Gardy**

Degree: **Doctor of Philosophy**

Title of Thesis: **Computational Prediction and Comparative Analysis of Protein Subcellular Localization in Bacteria**

Examining Committee:

Chair: **Dr. Erika Plettner**
Assistant Professor, Department of Chemistry

Dr. Fiona S.L. Brinkman
Senior Supervisor
Assistant Professor, Department of Molecular Biology and Biochemistry

Dr. Jamie K. Scott
Supervisor
Professor, Department of Molecular Biology and Biochemistry

Dr. Frederic Pio
Supervisor
Assistant Professor, Department of Molecular Biology and Biochemistry

Dr. Margo Moore
Internal Examiner
Professor, Department of Biological Sciences

Dr. Chris Upton
External Examiner
Associate Professor, Department of Biochemistry and Microbiology
University of Victoria

Date Defended/Approved: Friday, April 7, 2006

ABSTRACT

Predicting the subcellular localization of a protein is a critical step in processes ranging from genome annotation to drug and vaccine target discovery. Previously developed methods for localization prediction in bacteria exhibit poor predictive performance and are not conducive to the high-throughput analysis required in this era of genome-scale biological analysis. We therefore developed PSORTb, a high-precision, high-throughput tool for the prediction of bacterial protein localization. PSORTb implements a multi-component approach to prediction, incorporating the detection of several sequence features known to influence subcellular localization. With a reported overall precision of 96%, it is the most precise method available and one of the most comprehensive methods – capable of assigning a query protein to one or more of four Gram-positive or five Gram-negative localization sites. The PSORTb algorithm comprises a series of analytical steps, each step – or module – being an independent piece of software which scans the protein for the presence or absence of a particular sequence feature. Modules include: SCL-BLAST for homology-based detection, the HMMTOP transmembrane helix prediction tool, a signal peptide prediction tool, a series of frequent subsequence-based support vector machines, as well as motif and profile-matching modules. The modules return as output either a predicted localization site or – if the feature is not detected – a result of "unknown". The output is then integrated by a Bayesian network into a final prediction. Development of PSORTb also required the creation of PSORTdb, a database storing both known and predicted localization information for bacterial proteins. This is a valuable resource to both the localization prediction and microbial research communities, providing a source of training data for new predictive algorithms and acting as a discovery space. The release of PSORTb v.2.0 allowed us to carry out a number of analyses related to localization. We performed the first genome-wide computational and laboratory screen for N-terminal signal peptides in the opportunistic pathogen *Pseudomonas aeruginosa,* used PSORTb as a complement to laboratory-based high-throughput 2D gel studies of individual cellular compartments, and examined protein localization in a global context, revealing trends with implications for adaptive evolution in microbes.

Keywords:

Bioinformatics, localization, pathogenomics, proteomics, bacteria.

To my parents,

Terry Gardy and Sharon Harris,

with thanks for their never-ending love and support.

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks to everyone involved in the PSORTb project. First and foremost, Dr. Fiona S.L. Brinkman and the members of the Brinkman Lab – past and present – who collaborated on the PSORTb work: Michael Acab, Matthew Laird, Dr. Sébastien Rey, Cory Spencer, and Christopher Walsh. Thank you also to our collaborators in Simon Fraser University's Computing Science Department: Fei Chen, Dr. Martin Ester, Dr. Oliver Schulte, Rong She, and Dr. Ke Wang; at the University of British Columbia: Dr. Raphael Gottardo, Dr. Robert E.W. Hancock, and Dr. Shawn Lewenza; and around the world: Katalin deFays, Dr. Sujun Hua, Dr. Christophe Lambert, Dr. Istvan Simon, and Dr. Gabor Tusnády.

Special thanks to Dr. Kenta Nakai, who developed the first version of PSORT in 1991 and graciously allowed us access to his source code and offered much helpful advice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

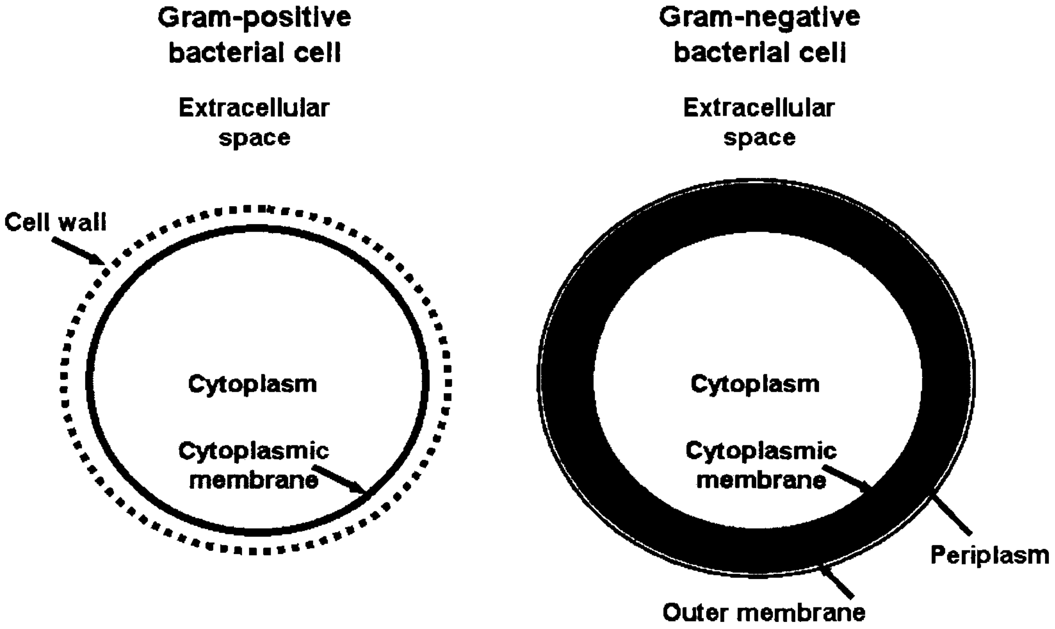
ABC	ATP-binding cassette
Beta version	initial non-public release of software, used for bug testing
BLAST	basic local alignment search tool
C	cytoplasm
CM	cytoplasmic membrane
CW	cell wall
EC	extracellular
Export	transport across the cytoplasmic membrane
FASTA format	a common sequence format: definition line followed by sequence
FN	false negative
FP	false positive
GFP	green fluorescent protein
GNU GPL	GNU general public licence (open source software licence)
GO	gene ontology
GSP	general secretory pathway
GST	generalized suffix tree
MinSup	minimum support
MySQL	structured query language: used in database systems
OM	outer membrane
P	periplasm
PhoA	alkaline phosphatase
Predictive coverage	% of proteins for which localizations are predicted
SDP	secreton dependent pathway
Secretion	transport out of the cell into the extracellular space
SRP	signal recognition particle
Subproteome	the proteins isolated from a single cellular compartment
SVM	support vector machine
TAT	twin arginine transporter
TMH	transmembrane helix
TN	true negative
TP	true positive
VLDC	variable length don't care

1 AN INTRODUCTION TO PROTEIN SUBCELLULAR LOCALIZATION IN BACTERIA

1.1 Protein localization in the bacterial cell

The Eubacterial domain of life comprises a diverse group of bacteria. Bacterial cells come in many shapes and sizes; however, regardless of the cell's gross morphology, the underlying structure of all bacterial cells can be defined quite simply. Microscopy, staining, and other analytical techniques have revealed that all bacterial cells consist of a cytoplasm – or cytosol –surrounded by a phospholipid bilayer– the cytoplasmic membrane. Beyond this innermost part of the cell, however, bacteria can be divided into two groups, each of which shows markedly different outer layers (Figure 1.1). These groups both differ from cells in the Archaebacterial domain, which, though similar in general appearance, do exhibit some differences at the physiochemical level.

Figure 1.1: The Gram-positive (left) and Gram-negative (right) bacterial cell.



The Gram-positive bacteria are the simpler of the two groups, surrounded by a thick layer of peptidoglycan – or murein – known as the cell wall. The Gram-negative bacteria are surrounded by a structure known as the cell envelope. This consists of a comparatively thin peptidoglycan layer – the equivalent of the Gram-positive cell wall, a second membrane called the outer membrane, and the space between the cytoplasmic – or inner – and outer membranes, termed the periplasm. This outer membrane is notably different from the symmetrical phospholipid bilayer forming the inner membrane – it is asymmetrical, with the outermost leaflet containing a molecule called lipopolysaccharide, or endotoxin.

All bacterial proteins are synthesized in the cytoplasm and indeed many remain here, however a number of proteins are targeted to one or more of the cellular compartments – or localization sites – described above. In both Gram-negative and Gram-positive bacteria, a protein may also be secreted out of the cell entirely into the extracellular environment.

Thus, in Gram-positive bacteria, a protein may be generally targeted to one or more of four localization sites: the cytoplasm, cytoplasmic membrane, cell wall, or extracellular space, while in Gram-negative bacteria, a protein may be targeted to one or more of five sites: the cytoplasm, cytoplasmic membrane, periplasm, outer membrane or extracellular space.

1.2 Signals governing protein targeting in bacteria

1.2.1 Bacterial transport systems

In order to carry out its function within the cell, a bacterial protein must frequently be targeted to a compartment other than the cytoplasm. This process necessitates traversing one or more localization sites, and is facilitated by the cell's complement of transport systems.

Bacterial transport systems are reasonably well-conserved between Gram-negative and Gram-positive bacteria, despite the difference in morphology between the two groups. The discussion below focuses on systems present in the more complex Gram-negative bacteria, with section 1.2.6 describing differences between the systems described and those present in Gram-positive organisms. A description of the targeting signals associated with each pathway is given in sections 1.2.2 – 1.2.5.

1.2.2 Type I transporters – ABC transporters

Type I transport is considered by some to be the simplest of the transport systems in bacteria – proteins are shuttled from the cytoplasm directly to the extracellular space in one step, using a multi-protein translocator called an ABC transporter which spans the entirety of the cell envelope (Holland et al., 2005). Energy for the process is produced by ATP hydrolysis, and the system is able to transport substrates ranging in size from 19kDa (Letoffe et al., 1994) to 800kDa (Hinsa et al., 2003). The canonical Type I transport system is the *E. coli* haemolysin system (Hly), consisting of the cytoplasmic membrane protein HlyB, the periplasm-spanning protein HlyD, and the outer membrane channel TolC (Koronakis and Hughes, 1993).

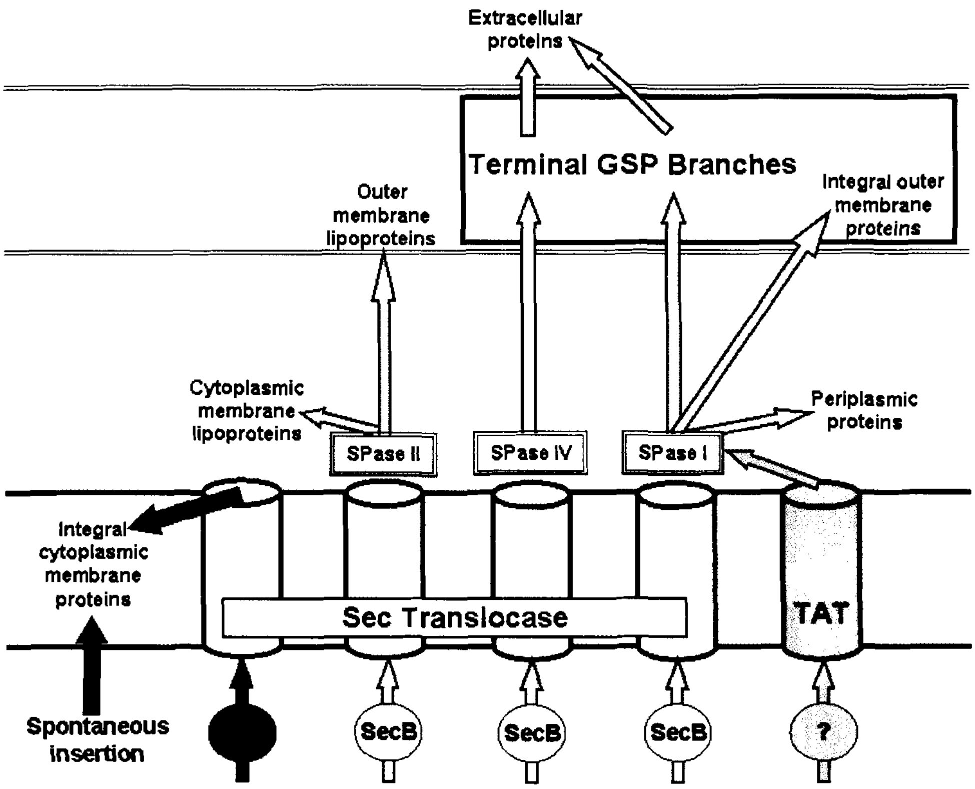
Proteins destined for secretion through the type I system are recognized post-translationally via a C-terminal signal, which is typically located in the vicinity of a glycine-rich repeat (Mackman et al., 1987; Delepelaire and Wandersman, 1990; Letoffe and Wandersman, 1992). With the exception of certain metalloproteases which contain a C-terminal DFVV motif (Ghigo and Wandersman, 1994), type I targeting signals are neither well-conserved nor defined by characteristic motifs.

1.2.3 Type II transporters – the general secretory pathway (GSP)

The general secretory pathway, or GSP, is the transport system used by the majority of exported proteins and is well-conserved between both Gram-negative and Gram-positive bacteria. In the first step of the GSP, proteins are targeted to and translocated across the cytoplasmic membrane, typically via the

Sec-dependent pathway. For Gram-positive bacterial proteins, this results in a membrane, cell wall or extracellular localization, while in Gram-negative organisms, this results in a cytoplasmic membrane or periplasmic localization. In the second step of the GSP, required only in Gram-negative bacteria, proteins are directed to one of multiple terminal branches of the pathway for targeting to the outer membrane or the extracellular space. Figure 1.2, based on a figure by Pugsley (1993a), presents a schematic representation of the GSP in Gram-negative bacteria.

Figure 1.2: The general secretory pathway in Gram-negative bacteria.



In the first step of the GSP, proteins are directed to and across the cytoplasmic membrane. For proteins whose final localization is the cytoplasmic membrane, this is either accomplished via spontaneous insertion of the protein into the membrane (de Gier et al., 1998), or through SRP (signal recognition particle)-mediated translocation. In the latter process, SRP – comprising a 4.5S RNA and the Ffh protein – recognizes hydrophobic segments of a nascent polypeptide (Luirink et al., 1992), and directs the protein to the FtsY receptor (Luirink et al., 1994). The interaction between Ffh and FtsY results in the release of the nascent protein from the SRP and its transfer to the Sec translocase.

Most proteins destined for other localization sites are post-translationally shuttled directly to the Sec translocase by the SecB chaperone. The exact region of protein recognized by SecB remains unknown, however all proteins utilizing this pathway exhibit N-terminal signal peptides and this leader sequence may thus represent the SecB substrate (Nakai, 2000). Following binding of SecB to a protein, SecB binds its receptor, SecA, a membrane-associated protein that provides the energy for the subsequent translocation step (Fekkes and Driessen, 1999). SecA then interacts with the Sec translocase, comprising SecY, SecE and SecG (Nishiyama et al., 1994).

A small proportion of Gram-negative bacterial proteins destined for the terminal branches of the GSP are exported by a Sec-independent pathway, the TAT – or twin-arginine – transporter. Although the chaperone responsible for recognizing folded proteins and directing them to the TAT transporter has not yet been elucidated, the structure of the transporter apparatus itself is known. It comprises three cytoplasmic membrane proteins: TatA, TatB and TatC. TatA represents the translocation pore, and TatB and TatC are thought to be involved in substrate recognition (Muller and Klosgen, 2005).

In the second step of the GSP, Gram-negative bacterial proteins are directed to the outer membrane or the extracellular space. In the case of outer membrane proteins, it is thought that the proteins insert into the outer membrane (de Cock et al., 1990a, 1990b) with the assistance of both a chaperone and binding to a periplasmic lipopolysaccharide (Kleinschmidt and Tamm, 2002). Extracellular proteins, however, must be directed to a terminal branch of the GSP for translocation across the outer membrane. Multiple terminal branches exist and tend to be specialized with regard to their substrates. The two well-studied branches are involved in the translocation of pili and of substrates including enzymes and toxins.

Pili proteins bind the pilin chaperone PapD in the periplasm (Hultgren et al., 1989) and subsequently bind the outer membrane PapC, through which they are translocated and assembled into a pilus (Dougan et al., 1983). Extracellular enzymes and toxins are translocated via what was initially referred to as the main terminal branch of the GSP, but which is now known as the secreton-dependent pathway (SDP). In this system, best represented by translocation of the *Klebsiella oxytoca* pullulanase (Pugsley, 1993b), a single substrate is translocated across the outer membrane by a dedicated secreton – a protein complex comprising 12-16 subunits (Sandkvist, 2001).

The signals directing a protein to the GSP determine which of the intial branches of the pathway a protein will take. Transmembrane alpha-helices appear to be sufficient to direct a protein to the SRP-mediated Sec pathway, while N-terminal signal peptides are required for Sec and TAT-mediated translocation. These are typically short sequences with a tripartite positive-hydrophobic-polar character suitable for partitioning into lipid bilayers. The N region is positively charged, the hydrophobic H region is a minimum of 8 hydrophobic residues in length and forms a membrane-spanning helix, and the C-region often contains a signal peptidase recognition site (von Heijne, 1985).

The majority of tripartite signal peptides are cleaved by signal peptidase I (SPase I), and are known as type I signal peptides. In some cases however, such signal peptides may remain uncleaved, serving as signal anchors for inner membrane proteins (von Heijne, 1988). Other types of tripartite signal peptides exhibit unique characteristics in their structure and are cleaved by different types of signal peptidases. For example, TAT substrates contain signal peptides similar to those processed by SPase I, however they also contain the twin arginine motif RRXFL[KR] upstream of their hydrophobic region (Chaddock et al., 1995). Signal peptidase II cleaves type II signal peptides, which are associated with lipoproteins. Although these are similar to type I signal peptides, cleavage occurs immediately upstream of a Cys residue, which is part of the N-terminal lipobox motif that characterizes these signal peptides (von Heijne, 1989). Prepilin peptidase cleaves type IV signal peptides, which differ from traditional N-terminal signal peptides in that they are short (~6 residues) with no tripartite structure. Instead cleavage occurs downstream from a glycine residue that precedes a long N-terminal stretch of hydrophobic amino acids in the mature protein (LaPointe and Taylor, 2000). A GFTLIE motif is often found in prepilin peptidase substrate signal peptides (Lory, 1994).

Signals directing a periplasmic intermediate to the outer membrane or to a terminal branch of the GSP are not well-understood, however. The proposed model of insertion for outer membrane proteins implies a structural basis for targeting, similar to the fashion in which transmembrane alpha helices direct the insertion of cytoplasmic membrane proteins. Outer membrane proteins adopt a beta-barrel structure, comprising an even number of beta strands – from eight (Vogt and Schulz, 1999) to 22 (Locher et al., 1998; Ferguson et al., 1998) – arranged in an anti-parallel fashion to form a barrel-like pore. Unlike the alpha-helices found in cytoplasmic membrane proteins, though, these beta-strands provide little in the way of information content (Schulz, 2002), beyond the fact that the C-terminal residue is frequently phenylalanine (Pohlner at al., 1987). Sequences of known strands exhibit little similarity to each other, and they tend to be about half the length of an alpha helical segment. Thus specific signals directing a protein to the terminal branches of the GSP have not yet been determined, although the role of periplasmic chaperones in the process is becoming increasingly apparent (Schleiff and Soll, 2005).

1.2.4 Type V transporters – autotransporters

Type V transporters are unique to the Gram-negative bacteria (Henderson et al., 1998). These proteins are autotransporters – or self-transporters – comprising an N-terminal passenger domain and a C-terminal transporter domain. The C-terminal domain forms a beta-barrel in the outer membrane through which the passenger domain is translocated. In most cases, the passenger domain is then cleaved and released into the extracellular milieu (Klauser et al., 1990, 1992). The translocation of *Neisseria gonorrhoeae* IgA1 is the canonical example of type V transport (Pohlner et al., 1987). Most autotransporters possess a type I signal peptide, however, as with the integral outer membrane proteins in the GSP system, little is known about the targeting signals contained within the transporter domain.

1.2.5 Type III and IV transporters – delivery of DNA/protein into host cells

The two remaining transport systems in bacteria are primarily involved in the direct injection of cytoplasmic DNA or protein substrates into host cells upon contact. In type III secretion, an effector protein is translocated from the cytoplasm into a eukaryotic cell through the needle complex, a multimer comprising as many as 20 different proteins (Hueck, 1998). Hypotheses regarding the signal that targets an effector for secretion through the needle complex include a motif in the 5' end of the effector mRNA (Anderson and Schneewind, 1999) and an N-terminal chaperone binding site located in the first 20 amino acids of the effector (Lloyd et al., 2001). Due to the markedly low sequence similarity between different effector proteins, and the lack of understanding regarding signals involved in Type III transport, identifying proteins localized by the Type III system has been notoriously difficult.

In type IV secretion, which remains poorly understood, DNA, nucleoproteins and proteins are translocated using machinery related to that involved in conjugation (Christie, 2001). Substrates are first translocated across the inner membrane using both Sec-dependent and Sec-independent pathways, and are then translocated across the outer membrane by an assembly of pilus-like proteins (Christie, 2001; Fischer et al., 2002; Burns, 2003). No signal targeting type IV effectors for secretion is yet known.

1.2.6 Differences in protein transport in Gram-positive bacteria

Gram-positive bacteria lack the outer membrane and periplasm of Gram-negative bacteria, and are instead surrounded by a thicker layer of peptidoglycan, termed the cell wall. With a significantly less complex cell wall, protein transport in this class of organisms is a simpler affair than in Gram-negative organisms, and most transport systems are not present. Protein targeting signals in Gram-positive bacteria, however, are quite similar to those of the Gram-negatives.

Most Gram-positive bacterial proteins are secreted via the Sec-dependent pathway, a pathway which is strongly conserved between Gram-negative and Gram-positive bacteria. Type I and II signal peptides, found in both classes of organism, are also remarkably similar in overall composition, however signal peptides from Gram-positive organisms do tend to be longer (Nielsen et al., 1997). Gram-positive bacteria also employ type I secretion via ABC transporters, however in these organisms, the transporter must span only the cytoplasmic membrane. In *B. subtilis*, the transporter can consist of one to four proteins. In the four-protein system, the transporter comprises two independent integral membrane proteins and two cytoplasmic substrate-binding proteins, while in the one-protein system, these components have fused into a single channel (Quentin et al., 1999).

1.3 Laboratory-based methods for localization determination

A protein's localization can be determined in the laboratory through any one of several techniques and their many variations. Only certain techniques are readily applicable to bacteria, however, and certain methods for localization determination are more common than others. The following section provides a brief introduction to some of the most frequently used techniques for determining protein subcellular localization for bacteria, along with a summary of their limitations.

1.3.1 Microscopy-based visualization

Microscopy-based methods for localization are employed quite frequently on account of their high precision and the quality of the information they provide – visualization of a protein *in situ* not only allows one to define its localization, but also permits one to examine its localization over time, as some proteins can be found in different subcellular compartments at different time points. In these methods, a protein of interest is tagged using a fluorescent protein or is incubated with a labelled antibody. Cells are then visualized under a microscope, with areas of fluorescence indicating the localization of the tagged or antibody-bound protein.

Green fluorescence protein (GFP), a 238 amino acid fluorophore from the jellyfish *Aequoria victoria*, is frequently fused to a protein of interest for such studies, as GFP itself is not specifically localized within the cell (Chalfie et al., 1994). Provided a successful fusion can be generated, the product of any gene can be visualized. In immunofluoresence microscopy, however, visualization is only possible if an antibody to the protein of interest is available. The antibody is coupled to a fluorescent dye, such as rhodamine or fluorescein, which permits visualization. While both of these methods typically use confocal scanning microscopy at the visualization stage, localization determination by electron microscopy is also possible when a protein has been tagged with an electron-dense particle such as colloidal gold, for example (Strachan and Read, 2004).

1.3.2 PhoA fusion

A second approach for localization determination in Gram-negative bacteria also relies on gene fusion, but in this case, visualization simply requires growth of a bacterial culture.

The enzyme alkaline phosphatase is encoded by the gene *phoA* and exhibits localization-dependent activity: the protein is active only when localized to the periplasm. In the alkaline phosphatase fusion technique (Manoil and Beckwith, 1985), the gene of interest is fused to a copy of *phoA* which has been truncated to remove its native type I signal peptide. If the product of the gene of interest contains an export signal, the *phoA*:gene fusion product will reach the periplasm and become enzymatically active. When grown on 5-bromo-4-chloro-3-indolyl phosphate, colonies containing the fusion product will appear blue. If the gene of interest's product does not contain an export signal, the fusion product will remain in the cytoplasm and growth on on 5-bromo-4-chloro-3-indolyl phosphate will yield white colonies.

Using the *phoA* fusion technique, proteins containing an export signal directing them to the periplasm and beyond can be identified. This includes proteins with signal peptides, transmembrane alpha-helices, or other sequence-encoded targeting motifs.

1.3.3 Subcellular fractionation

Perhaps the most classic technique for determining protein localization is subcellular fractionation followed by protein identification (Albertsson, 1956). In this approach, the bacterial cell is first separated into its constituent compartments using a series of detergent extractions and/or centrifugations. The proteins resident in each compartment are then resolved using one of several possible methods, including gel electrophoresis, chromatographic separation, and/or mass spectrometry. This technique not only provides immediate evidence of a protein's localization, but it also has the advantage of generating results for a large number of proteins in a single analysis.

1.3.4 Limitations of laboratory-based methods

The methods described above and other laboratory-based techniques are capable of experimentally verifying a protein's localization. This does not mean, however, that they always produce a correct result. False positives and negatives are a possibility in any of these analyses, and each exhibits its own specific limitations.

In the case of fluorescently tagged proteins, tag insertion has the potential to disrupt a targeting sequence, resulting in an improperly localized gene product. This is especially true for proteins utilizing a C-terminal secretion signal. Tagged proteins also display a tendency to aggregate. Microscopy-based methods are also difficult to perform in a high-throughput fashion. Production of a large number of gene fusions does not yield a 100% success rate, and screening of successful fusions requires automated visualization methods which may not always be correct in their assessment of localization (Hu and Murphy, 2004).

PhoA fusions are only possible in Gram-negative bacteria, and cannot provide information beyond whether a protein is exported to – or past – the periplasm. Like GFP fusions, PhoA fusions can disrupt a native targeting signal, resulting in a false negative, and any fusion experiment runs the risk of cytotoxicity.

Fractionation studies, while powerful, are limited to the identification of those proteins expressed in a cell at a specific time under specific conditions, thus determining the localization of a protein of interest may not always be possible. Low-abundance proteins can also be easily missed (Stasyk and Huber, 2004). Furthermore, the identification of the proteins isolated is not always possible – sometimes a database match to a protein cannot be determined. Contamination by proteins from neighbouring subcellular fractions may also occur.

Most importantly, regardless of the method used, there are two basic limitations which cannot be overcome in the lab. Laboratory-based localization determination methods require a notable investment of time and laboratory resources in comparison to a computational prediction, which can be generated quickly using a computer alone.

1.4 Computational methods for localization prediction in bacteria

1.4.1 The importance of computational predictive methods

As described above, laboratory-based methods for protein localization carry a number of caveats, chief among them the time and resources required. In order to reduce the investment of effort and money that go into a localization experiment, it is desirable to narrow the focus of such an experiment using pre-existing knowledge. By computationally identifying one or more signals known to influence or correlate with protein localization, a prediction of which cellular compartment, or compartments, a protein is likely resident at can be generated using sequence information alone. With the explosion in publicly available sequence data, such prediction has become a critical component of biological research. Indeed, without the rapid, high-throughput information provided by computational analysis of sequences in general, the large amounts of data generated by sequencing projects cannot be used to their full potential. Predicted protein localization data, in particular, affords a number of insights that can aid in the prioritization of proteins for further study.

Predicted localization information can be used in the genome annotation process. The ultimate goal of many annotation projects is to generate predicted functions for each gene product in the genome. Annotation transfer by homology is typically used to assign function to proteins, however this is only possible for the portion of an organism's gene complement that shows similarity to other annotated genes. For the remaining proteins, other functional clues are required. Because cellular compartment and function are closely related, a protein's predicted localization can provide clues to its function and vice-versa; for example, integral outer membrane proteins are typically involved in the uptake and/or efflux of specific substrates, while a protein annotated as DNA-binding is likely to be found in the cytoplasm.

Knowing the cellular compartment that a protein is likely resident at can also aid in experimental design and proteomics-based analysis. If an attempt to isolate a particular protein is being made, predicting the protein's localization can narrow down the search space significantly. Rather than inspecting a whole cell lysate for a protein of interest, for example, the cell can be fractionated into its constituent compartments and only the compartment of interest analyzed. Section 7 describes how predicted localization information can also be employed to screen the results of a subfractionation analysis for potential contaminants or other errors.

Of all the potential applications of localization prediction, perhaps most relevant to the medical community is the fact that being able to rapidly identify the surface-exposed proteins in a bacterial genome can facilitate the discovery of novel drug targets and potential vaccine components. Traditionally, "subunit" vaccines against bacterial infection have been formulated by infecting an animal with a bacterium, using antisera to identify surface-exposed immuno-reactive proteins, purifying these proteins and cloning them into an expression system, and then analyzing individual regions of the cloned protein to identify especially reactive protein subunits for use as vaccine candidates (Chakravarti et al., 2000). Localization prediction by computer, however, could potentially identify all of the potentially surface-exposed proteins encoded in a given genome within minutes, significantly narrowing the search space for drug and vaccine target discovery and reducing the time and expense associated with these procedures. There is therefore a strong interest in improving the computational prediction of protein subcellular localization for medically-relevant bacteria. In addition, such subcellular localization prediction can also aid in the identification of cell-surface proteins that may be suitable targets for a new diagnostic/detection method (Ertl et al., 2003). For non-pathogenic bacteria of environmental importance, there is also an interest in identifying cell-surface proteins as part of efforts to develop microbial detection methods to identify such microbes in environmental samples.

1.4.2 Early computational methods for localization prediction in bacteria

The roots of computational prediction of protein localization lie in the identification of individual sequence features known to influence or correlate with localization. Many of the first approaches involved the prediction of type I signal peptides – N-terminal protein sequences directing the export of a protein out of the bacterial cytoplasm via the Sec machinery. Weight matrix-based analyses or related approaches were frequently employed to predict these signal sequences. In this technique, frequency values reflecting each amino acid's occurrence in known signal peptides are assigned to each residue in a query sequence. A sliding window then moves down the sequence, summing the frequency scores. In this fashion, the region most likely to represent a signal peptide can easily be identified (McGeoch, 1985; von Heijne, 1986; Folz and Gordon, 1987; Popowicz and Dash, 1988). Neural network techniques (Ladunga et al., 1991; Schneider et al., 1993; Schneider and Wrede, 1993; Nielsen et al., 1997) and Hidden Markov Models (Nielsen and Krogh, 1998) were later employed, offering an improvement in predictive power.

Other approaches to localization prediction involved the prediction of transmembrane alpha helices – secondary structure elements that traverse the cytoplasmic membrane. The first of these methods implemented sliding windows and hydrophobicity scales to search for membrane spanning segments (Kyte and Doolittle, 1982; Eisenberg et al., 1984), while later methods introduced additional considerations, including the use of the "positive-inside rule", which states that positively charged residues occur at a higher frequency on the cytoplasmic face of the membrane (von Heijne, 1992), and the use of neural networks and Hidden Markov Models (Nakai and Kanehisa, 1992; Hofman and Stoffel, 1992; Claros and von Heijne, 1994; Jones et al., 1994; Rost et al., 1996; Cserzo et al., 1997; Persson and Argos, 1997; Sonnhammer et al., 1998; Tusnády and Simon, 1998; White and Wimley., 1999; Deber et al., 2001; Krogh et al., 2001; Juretic et al., 2002).

While signal peptide and transmembrane helix predictions represented a critical first step on the route to complete localization prediction, their limited utility is clear. Each method is only capable of providing information regarding a single localization site and, in the case of a positively-identified signal peptide, the extent of this information is only the knowledge that the protein is likely not cytoplasmic. False positive results were a frequent problem, and false negatives an even greater problem. Early signal peptide prediction methods in particular were not capable of recognizing a variety of non-traditional sorting signals.

At the same time as great strides were being made into feature prediction as described above, early work into computationally sorting a protein to one of multiple localization sites was beginning. In 1991, Nakai and Kanehisa released PSORT I, an expert system for localization prediction in Gram-negative bacteria. Capable of sorting a query protein to the cytoplasm, cytoplasmic membrane, periplasm or outer membrane (but not the extracellular space), PSORT I represented the first true protein subcellular localization prediction method. The program employed a multi-component approach to prediction: features influencing localization – including amino acid composition, signal peptides, functional motifs and transmembrane helices – were identified in a query protein, and the resulting information was integrated to generate a final prediction using an "if-then" rule system. On a set of 106 bacterial proteins, PSORT I was able to assign 83% of them to the correct localization site. The method was updated in 1999, replacing the if-then expert system with the *k*-nearest neighbour algorithm and improving the algorithm's reasoning slightly (Nakai and Horton, 1999).

No further localization prediction methods were developed until 1997, when Cedano et al. utilized the differences in amino acid composition between proteins resident at different cellular compartments as the basis for a sorting algorithm. Their method, ProtLock, was designed to sort eukaryotic proteins, and correctly predicted the localization of 76% of the 200 training proteins. Andrade et al. continued this work in 1998, again using eukaryotic proteins.

In 1998, Reinhardt and Hubbard exploited the differences in amino acid composition to create a neural network capable of sorting a bacterial query protein to one of three sites, the cytoplasm, periplasm, or extracellular space. This tool, NNPSL, achieved a prediction accuracy of 81%. Chou and collaborators also developed a series of tools employing the discriminant function (Chou and Elrod, 1998; Chou and Elrod, 1999; Chou, 2000), neural networks (Cai et al., 2002), and support vector machine (SVM) (Cai et al., 2000) to analyze amino acid composition in bacteria, however none of the resulting software was released publicly.

For over a decade, PSORT I remained the predominant computational method used by researchers to make subcellular localization predictions for bacterial proteins. Factors contributing to its widespread use include the fact that the tool was the first of its kind to be developed, that for over a half a decade following its release it represented the only available method, and it was freely accessible over the internet. However, in the years following PSORT I's release there were considerable improvements in bioinformatics algorithm development in general, as well as a rapid expansion of knowledge regarding protein sorting signals. For this reason we undertook the challenge of developing a new, comprehensive subcellular localization predictor for bacterial proteins, using these updated computational methods and our expanded biological knowledge.

1.4.3 Recent computational methods for localization prediction in bacteria

This thesis describes the creation of PSORTb, a high-precision, high-throughput open-source tool for the prediction of bacterial protein localization. Over the course of PSORTb's development, a number of other localization prediction tools were also released. The underlying principle and availability of each of these methods is summarized in Table 1.1. Section 5 presents a comparison of the predictive performance of these methods relative to PSORTb.

Table 1.1: A summary of available computational methods for bacterial protein localization prediction.

Program	Reference	Analytical method	Localizations predicted	Usage	Open source	Forces output?
PSORT I	Nakai and Kanehisa, 1991	Multi-component	4 Gram-negative 3 Gram-positive	Web Local (Sun/Solaris systems)	No specified licence	No
PSORTb	Gardy et al., 2003, 2005	Multi-component	5 Gram-negative 4 Gram-positive	Web Local (Most UNIX/Linux systems)	GNU General Public Licence	No
Proteome Analyst	Lu et al., 2004	Annotation keywords	5 Gram-negative 3 Gram-positive	Web	No	No
SubLoc	Hua and Sun, 2001	SVM	3 (no Gram distinction)	Web	No	Yes
CELLO	Yu et al., 2004	SVM	5 Gram-negative 4 Gram-positive	Web	No	Yes
PSLpred	Bhasin et al., 2005	SVM	5 Gram-negative	Web	No	Yes
LOCtree	Nair and Rost, 2005	SVM	3 (no Gram distinction)	Web	No	Yes
P-CLASSIFIER	Wang et al., 2005	SVM	5 Gram-negative	Web	No	Yes

1.4.3.1 Proteome Analyst – a keyword-based approach

Proteome Analyst's subcellular localization prediction server (Lu et al., 2004) employs an annotation keyword-based approach comprising two steps. In the first, a query protein is compared, using BLAST, against the SwissProt database, returning a set of homologs with manually curated annotation. Keywords in the annotation that might be indicative of a particular localization site are extracted from the SwissProt records and, in the second step, are passed to a Naïve Bayes classifier specific to the class of organism. This classifier then uses the extracted keywords to assign the query protein to one of three Gram-positive or five Gram-negative localization sites. Proteome Analyst returns a final prediction and an associated confidence score on the 100.0% scale. The program is also capable of generating predictions for animal, plant, and fungal sequences.

Proteome Analyst can be accessed at http://www.cs.ualberta.ca/~bioinfo/PA/Sub/, with the server accepting single or multiple sequences as input. A choice of two output formats is provided – detailed HTML output or a shorter format comma separated value file. BLAST results for the query protein are also provided. While the tool is not available for download and thus cannot be used locally, the site does host the PA-GOSUB database, containing a selection of precomputed predictions for microbial and other genomes (Lu et al., 2005). The authors also note that they are willing to run predictions for specific genomes requested by users.

1.4.3.2 Amino acid composition support vector machine-based methods

SVM is a machine-learning technique frequently employed to solve binary classification problems (see section 4.3). Composition-based SVMs exploit the differences in frequency of the 20 amino acids across different cellular compartments, and are thus capable of making predictions when no prior information about a protein is available, such as homologs in existing databases or predicted sequence features.

SubLoc (Hua and Sun, 2001) was the first publicly available SVM-based localization tool to be released. Capable of sorting bacterial proteins to the cytoplasm, periplasm or extracellular space, it generates predictions using a single SVM analyzing a protein's overall amino acid composition. The program returns a final prediction as well as two measures of predictive confidence – a reliability index score and an estimate of accuracy.

While the release of SubLoc marked an important milestone in the use of SVMs for localization prediction, the method itself carries two significant caveats. Because the program only sorts proteins to three compartments, known and suspected cytoplasmic membrane and outer membrane proteins must be removed prior to the analysis. Furthermore, the method does not distinguish between Gram-positive and Gram-negative queries, thus proteins from a Gram-positive organism can be mistakenly classified as periplasmic. SubLoc accepts web-based submissions of one or more sequences at http://www.bioinfo.tsinghua.edu.cn/SubLoc/ and returns output in the format of an

HTML table. The program is not available for download and use on a local computer – it must be accessed over the web.

CELLO (Yu et al., 2004) employs an extended version of SubLoc's composition-based SVM in which five SVMs are used – one analyzing overall amino acid composition, one incorporating sequence order information, and three utilizing modified composition analysis in which amino acids are grouped according to their physiochemical properties. The output of each SVM is integrated to generate a final prediction, which is returned along with a score distribution on a five-point scale. The program can assign a protein to one of five Gram-negative or four Gram-positive localization sites. CELLO accepts web-based submissions of one or more sequences at http://cello.life.nctu.edu.tw/ and returns output in a text format. The program is not available for local use.

Like CELLO, PSLpred (Bhasin et al., 2005) uses multiple SVMs to assign a query protein to one of five Gram-negative localization sites. However, in addition to three SVMs analyzing overall composition, dipeptide composition, and composition incorporating physiochemical groupings, PSLpred also implements a PSI-BLAST module for similarity searching. Output is returned in HTML format, which includes a final prediction, a reliability index on a five-point scale, and an estimate of accuracy. The tool is available at http://www.imtech.res.in/raghava/pslpred/, however registration is required to use the program and proteins must be submitted one at a time. The program is not available for local use.

LOCtree (Nair and Rost, 2005) combines SVM-based analysis with a flowchart-style decision system designed to mimic cellular sorting. In the first step, an amino acid composition-based SVM determines whether the query protein is cytoplasmic or non-cytoplasmic. Non-cytoplasmic proteins are then passed to a second SVM, which determines whether the protein is periplasmic or extracellular. Like SubLoc, LOCtree only assigns a protein to one of three localizations, thus membrane proteins must be removed from the dataset prior to analysis. Output is in the form of an HTML table, including a final prediction as well as a reliability index on a ten-point scale. The results of other analyses, including signal peptide prediction, secondary structure prediction, and motif searching are also provided. Up to 100 sequences at a time can be submitted to LOCtree at http://cubic.bioc.columbia.edu/cgi-bin/var/nair/loctree/query. The program is not available for local use.

P-CLASSIFIER implements 15 SVMs in its analysis, representing a combinatorial approach in which sequence fragments of length *n*, where *n* = 1-4, are examined and different physiochemical-based groupings of similar amino acids are employed (Wang et al., 2005). The method is capable of assigning a Gram-negative protein to one of five localization sites, and returns its output in an HTML table. Predictions are reported along with a distribution of scores using a percentage system. P-CLASSIFIER accepts web-based submissions of up to 100 sequences at a time at http://protein.bii.a-star.edu.sg/localization/gram-negative/introduction.html. The program is not available for local use.

1.5 Goal of the present research

At the outset of PSORTb development in 2001, only two tools – PSORT I and NNPSL – were available for the prediction of protein localization in bacteria. NNPSL was limited by the fact that it only recognized three localization sites, ignoring membrane proteins, while PSORT I was hampered by not recognizing proteins secreted to the extracellular space. Despite these limitations, the PSORT I program remained in widespread use in the microbiology community – in particular for early genome-wide analyses in the 1990's, as a freely available Sun/Solaris UNIX version for local use was made available.

Recognizing that there was a need for improvement, we set a goal of developing an improved, high-precision, high-throughput localization prediction tool for bacteria. Herein, I describe the development of this method – PSORTb – and the numerous applications we have found for the method, which have implications for the field of proteomics and for concepts underlying microbial adaptive evolution.

2 ePSORTdb: A DATABASE OF BACTERIAL PROTEINS OF KNOWN LOCALIZATION

Portions of this chapter have been previously published in the article "PSORTdb: a protein subcellular localization database for bacteria", co-authored by S. Rey, M. Acab, J.L. Gardy, M.R. Laird, K. deFays, C. Lambert and F.S.L. Brinkman in Nucleic Acids Research, Volume 33, Database Issue. © 2005 Oxford University Press.

2.1 Summary

ePSORTdb (http://db.psort.org) is a web-accessible database of protein localization data for bacteria that contains information determined through laboratory experimentation. The database, which contains approximately 2000 proteins, is manually curated and represents the largest dataset of its kind. ePSORTdb has been used for training localization prediction tools, including PSORTb, and represents an important resource for the localization prediction and microbiology communities. ePSORTdb can be accessed through the web using a very flexible text search engine, a data browser, or using BLAST, and the entire database or search results may be downloaded in various formats. Features such as GO ontologies and multiple accession numbers are incorporated to facilitate integration with other bioinformatics resources. ePSORTdb is freely available under the GNU General Public License.

2.2 The need for high quality training data

By definition, a machine-learning method requires a set of training data – known instances of the class of object the method is designed to predict or classify. For a bacterial protein subcellular localization predictor, these training data must thus comprise proteins whose localization within the cell has been experimentally verified. Because of the fundamental differences in cellular ultrastructure between bacteria and eukaryotes, the data must be bacterial in origin.

Prior to the development of our training dataset, most localization prediction methods – both bacterial and eukaryotic – were trained on data extracted from the SWISS-PROT database (Bairoch and Boeckmann, 1991, 1992, 1993, 1994; Boeckmann et al., 2003), in particular the data described by Reinhardt and Hubbard (1998). In a typical extraction procedure, all full-length, non-ambiguous protein sequences from the taxa of interest are screened to remove those lacking an annotation in the "subcellular location" field. The resulting list of proteins is then further filtered to remove sequences whose localization annotation contained the terms "by similarity" or "probable", implying that experimental confirmation of localization has not been performed.

While SWISS-PROT represents an excellent resource, it is not specifically designed with protein subcellular localization annotation in mind. As a result, several aspects of SWISS-PROT's design have potentially serious implications for training of a machine-learning method. First, heterogeneity among annotations is quite common, with multiple phrases often used to describe the same cellular compartment. While some of these distinctions are easily resolved, for example the use of both "cytosol" and "cytoplasm" to refer to the interior of a cell, others, such as "membrane-bound" versus "membrane-associated", can easily be misinterpreted. Second, many annotations lack sufficient information to conclusively assign a localization site. Many Gram-negative bacterial proteins are annotated simply as “membrane”, with no distinction being made between the cytoplasmic and outer membranes, while other membrane-associated proteins are not further annotated to indicate on which side of the membrane the bulk of the protein resides. Finally, many proteins are known to contain domains residing in two or more different cellular compartments, information that is rarely captured in a SWISS-PROT annotation. Because of these limitations, we chose to adopt an alternative approach to dataset development incorporating literature-derived evidence and manual review.

2.3 ePSORTdb development

Bacterial proteins with annotated localization information were extracted out of SWISS-PROT, and any sequence whose annotation was noted as “potential”, “probable”, or “by similarity” was discarded. The resulting dataset was then subjected to manual review. PubMed abstracts and full-text articles were searched using keywords from the SWISS-PROT entry in an attempt to find experimental verification of the annotated localization site. During this literature review process, it was observed that some proteins whose localization was reported in high-throughput proteomics manuscripts – primarily studies involving fractionation followed by 2D gel electrophoresis – had been erroneously annotated. Critical review of these studies revealed errors, including contamination of a particular cellular fraction by proteins known to be resident in other compartments. For this reason, proteins whose localization had been determined by high-throughput proteomics analyses were not included in ePSORTdb.

This extraction procedure followed by literature review was initially performed with release 39 of SWISS-PROT, giving rise to ePSORTdb v.1.0 and v.1.1, and was later updated using release 40.29, yielding the ePSORTdb v.2.0 dataset. To further expand ePSORTdb v.2.0, alternative literature sources were employed, including microbiology textbooks (Neidhardt et al., 1996; Fischetti et al., 2000; Sonenshein et al., 2001). Versions 1.0 and 1.1 of the dataset include Gram-negative data only, while version 2.0 is expanded to include Gram-positive data.

ePSORTdb recognizes both single and multiple bacterial localization sites. In Gram-negative bacteria, five single sites are included: the cytoplasm (C), cytoplasmic membrane (CM), periplasm (P), outer membrane (OM) and extracellular space (EC). Four multiple localization sites are also noted: C/CM, CM/P, P/OM and OM/EC. In Gram-positive bacteria, four single sites are included: the cytoplasm (C), cytoplasmic membrane (CM), cell wall (CW) and extracellular (EC); and two multiple localization sites: C/CM and CM/CW. Each single localization site term is associated with a unique Gene Ontology (GO) (Ashburner et al., 2000) identifier, while multiple localization sites are represented by a combination of their GO identifiers.

Experimentally verified localization sites contained in ePSORTdb are accessible in three formats: a terse, machine-readable definition (*e.g.*

cytoplasmic membrane), the associated GO identifier (*e.g.* 0005886), and a verbose definition (*e.g.* cytoplasmic membrane integral membrane protein).

2.4 ePSORTdb releases

Table 2.1 describes the composition of each of the three releases of ePSORTdb. Version 1.0 is described in Gardy et al. (2003), version 1.1 represents a small update made between publications, and version 2.0 is described in Gardy et al. (2005) and Rey et al. (2005a).

Table 2.1: Composition of ePSORTdb releases.

	Single Localization Sites					
ePSORTdb Release	**C**	**CM**	**CW**	**P**	**OM**	**EC**
Gram-negative v.1.0	248	268	-	244	352	190
Gram-negative v.1.1	278	292	-	276	377	191
Gram-negative v.2.0	278	309	-	276	391	190
Gram-positive v.2.0	194	103	61	-	-	181
	Multiple Localization Sites					
ePSORTdb Release	**C/CM**	**CM/P**	**P/OM**	**OM/EC**	**CM/CW**	
Gram-negative v.1.0	14	49	-	76	-	
Gram-negative v.1.1	16	51	2	78	-	
Gram-negative v.2.0	16	51	2	78	-	
Gram-positive v.2.0	15	-	-	-	20	

2.5 ePSORTdb database and website

As part of our PSORTb resource development, we decided to create a publicly accessible, flexible database to house our datasets of proteins of experimentally determined subcellular localization. Each ePSORTdb record contains extensive information regarding the sequence. NCBI's GI number (Benson et al., 2000) is used as the primary identifier, facilitating linkage to other databases. When available, additional fields relevant to protein identification are also included to facilitate searching and linking out to external resources: protein name, gene name, alternate protein and gene names, and SWISS-PROT accession number. Other fields further define the sequence in a broader sense: source organism name, phylum, class, NCBI taxonomy identifier (Wheeler et al., 2000), and Gram stain class. Links to the source of the annotation, in the form of a PubMed ID, book title, ISBN number or URL, are provided for some sequences. Finally, amino acid sequences and their length are also made available.

PSORTdb's data are housed in a MySQL database. Using PHP and JavaScript, the web database application – freely accessible at http://db.psort.org – was developed to facilitate access to the data without prior knowledge of SQL, relational databases or specifics of the ePSORTdb database schema. The browsing and dynamic textbox features of the web interface also make it easier for a user to search the data, even if one is unfamiliar with how the data are stored. Three search tools provide an entry point to the dataset.

With the text search tool, one or more keywords or other values suitable for a given field can be used to query the database against one or more data fields. Boolean operators are available to enable complex queries. A dynamic textbox displays a description and/or example of the type of text permitted for a particular search field. This feature assists users in choosing their queries. For example, if a user wishes to search the localization field, all possible localizations are presented in a dynamic textbox from which the user can select his or her terms of interest. This ensures the correct query is implemented and prevents common errors in query-based searching, including spelling mistakes or improper use of terminology.

The browse tool allows the user to explore the dataset in a hierarchical fashion similar to browsing the NCBI Taxonomy Database (Wheeler at al., 2000) or Gene Ontologies (Ashburner et al., 2000). The text used to populate the browsing function is dynamically generated from the MySQL database, and permits exploration of the data by localization, phylum, class, Gram stain and organism in every possible logical combination.

Sequences in FASTA format may be also submitted and searched against the database using a BLAST search (Altschul et al., 1997), which searches against a file system rather than the MySQL database. Results are returned in standard BLAST HTML format.

The text search and browse tools produce an HTML table of results that can be viewed page by page. Initially, a default set of fields is displayed; however there are numerous options available that allow a user to customize the display of results. A user can change the number of records viewed per page, simultaneously sort the table on up to three fields in ascending or descending order, select the fields to be displayed, and rearrange the order of the fields. In addition to viewing the results as an HTML document on a web browser, the user may also download the data as a tab-delimited file or a FASTA formatted file. From both the text search/browse result lists and the BLAST output, a user can click on a protein's GI number to obtain detailed annotations as described above.

A web form is also available through which researchers can submit proposed updates or corrections to the database, all of which are subject to manual review. This is an important component of the database, enabling researchers' participation and inclusion of their data, and the submission form has been made as simple as possible to encourage participation.

2.6 Applications of ePSORTdb

By providing a centralized, freely available localization resource which can be queried to isolate specific subsets of interest, ePSORTdb represents an important source of training data for researchers wishing to develop novel classification methods. Beyond its use as a source of training data, however, the potential applications of ePSORTdb are numerous in both the bioinformatics arena and in related fields. By implementing data mining or pattern discovery tools, a bioinformatician can discover features representative of a particular localization site quite readily. Microbiologists can employ the dataset in the identification of targets in bacterial genomes for surface proteins for environmental diagnostics, medical diagnostics, vaccines, antimicrobial compounds and other uses. Furthermore, the information contained in the dataset can also assist in the annotation of newly sequenced bacterial genomes. Proteomics researchers can utilize the data as a check in subfractionation experiments, and the information can also assist in experimental design.

3 DEVELOPMENT AND RELEASE OF PSORTb V.1.0

Portions of this chapter have been previously published in the article "PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria", co-authored by J.L. Gardy, C. Spencer,K. Wang, M. Ester, G.E. Tusnády, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F.S.L. Brinkman in Nucleic Acids Research, Volume 31, Issue 13. © 2003 Oxford University Press.

3.1 Summary

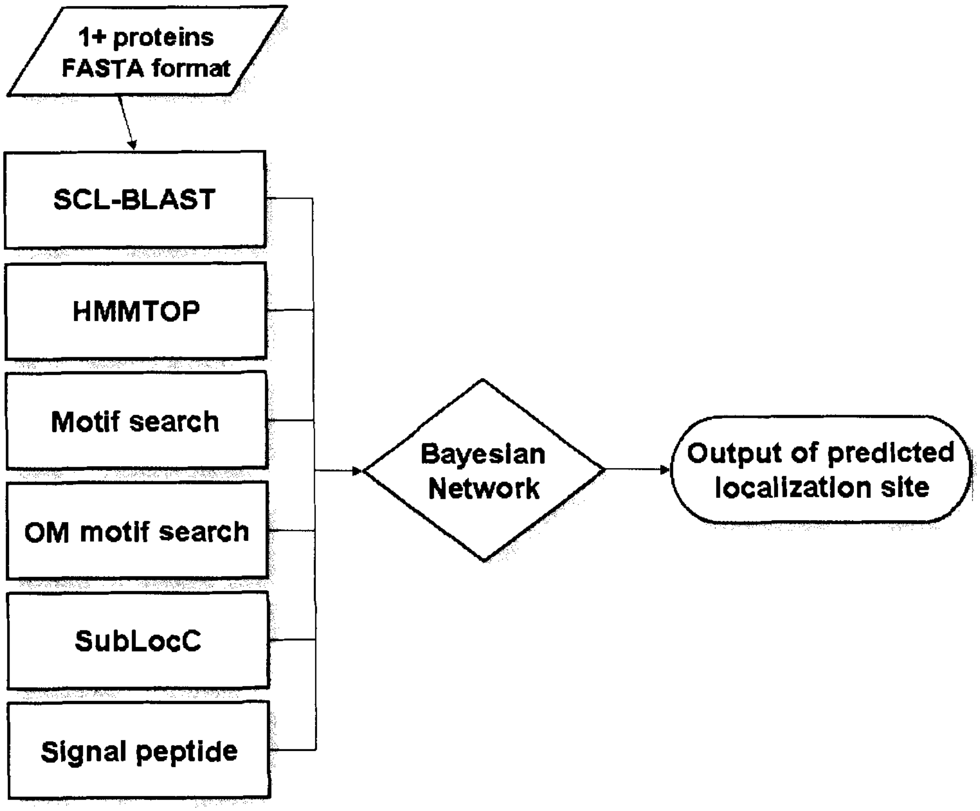
Automated prediction of bacterial protein subcellular localization is an important tool for genome annotation and drug discovery. PSORT I was one of the most widely used computational methods for such bacterial protein analysis from 1991 onwards; however, it had not been significantly updated since it was introduced in 1991. In addition, neither PSORT I, nor any of the other computational methods available at the outset of PSORTb development made predictions for all five of the localization sites characteristic of Gram-negative bacteria. We therefore developed PSORTb, an updated version of PSORT for Gram-negative bacteria, which was made available as a web-based application at http://www.psort.org. PSORTb examines a given protein sequence for amino acid composition, similarity to proteins of known localization, presence of a signal peptide, transmembrane alpha-helices and motifs corresponding to specific localizations. A probabilistic method integrates these analyses, returning a list of five possible localization sites with associated probability scores. PSORTb, designed to favor high precision (specificity) over high recall (sensitivity), attained an overall precision of 97% and recall of 75% in 5-fold cross-validation tests,

using a dataset we developed of 1443 proteins of experimentally known localization. The PSORTb source code is freely available under the GNU General Public License.

3.2 Predictive modules

The creation of the first version of the ePSORTdb dataset was followed by the development of the predictive modules that would form the initial release of PSORTb, a bacterial protein localization predictive tool designed for Gram-negative organisms. We hypothesized that the computational identification of sequence features known to influence or correlate with protein localization could be used to generate predicted localization site information for a query protein. A review of the literature identified several candidate features, and short software programs – termed modules – were developed to identify these features in an amino acid sequence. Modules include: SCL-BLAST, HMMTOP, motif searching, SubLocC, and a signal peptide identification tool (Figure 3.1).

Figure 3.1: Organization of PSORTb v.1.0.



3.2.1 SCL-BLAST

Subcellular localization tends to be evolutionarily conserved (Nair and Rost, 2002b), thus homology to a protein of known localization appears to be a good indicator of a protein's actual localization site. We therefore constructed a module entitled SCL-BLAST (for SubCellular Localization BLAST), in which a BLAST search (Altschul et al., 1997) of a submitted protein is carried out against ePSORTdb v.1.0, using an E-value cutoff of 10 e^{-10}. A length restriction is placed on resulting high scoring pairs, such that the length of the high scoring pair must be within 80–120% of the length of the subject. This reduces the potential for misprediction of localization based on similarity to a single domain of a protein in the database, a protein whose domains may reside in different localization sites.

The module returns the localization site and SWISS-PROT accession number of any hits fulfilling the above criteria and can generate a prediction for any of the five sites.

3.2.2 HMMTOP

Integral inner membrane proteins are characterized by the presence of alpha-helical transmembrane regions (von Heijne, 1994) and this feature has been used as a reliable indicator of localization at the inner membrane in past predictors, including PSORT I (Nakai and Kanehisa, 1991). PSORTb utilizes the Hidden Markov Model-based method HMMTOP (Tusnády and Simon, 1998, 2001) to identify potential transmembrane alpha helices, assigning a localization of inner membrane if three or more helices are found.

3.2.3 Motifs

A protein's functional description is often indicative of its subcellular localization (Eisenhaber and Bork, 1998). Therefore, certain sequence patterns corresponding to function may also correlate with a specific subcellular localization. PROSITE release 17.0 (Hofmann et al., 1999) was searched for such potential patterns and the resulting list was tested on ePSORTdb. Twenty six motifs, available at http://www.psort.org/motifs and in Appendix A, capable of identifying subcellular localization with 100% precision were retained. The module returns the localization site and PROSITE accession number of any pattern found within the sequence and can generate a prediction for any of the five sites.

3.2.4 Outer membrane protein motifs

The identification of outer membrane proteins is of particular interest, both due to the difficulty in predicting their characteristic beta-barrel structure and their high potential for use as drug targets. A data mining approach called association rule mining was used to identify frequent sequences occurring only in beta-barrel proteins – both integral outer membrane proteins and autotransporter proteins – which possess a beta-barrel transport domain (She et al., 2003). In this technique, "association rules" are computationally identified – these are attributes that are characteristic of a particular dataset. In our case, each attribute is a short amino acid subsequence that is characteristic of beta-barrel proteins. By varying the required support (how many beta-barrel proteins is the subsequence found in?) and confidence (is the subsequence likely to be found in a non-beta-barrel protein, giving a false positive?), we can identify the set of association rules, or subsequences, that provide the best classification of outer membrane vs. non-outer membrane proteins.

A total of 279 frequent sequences (Appendix B) were generated and used to build a classifier. A user-submitted sequence is screened for the presence of three or more of the frequent sequences and is classified as either outer membrane or non-outer membrane based on the result.

3.2.5 SubLocC

Support Vector Machine, or SVM, has been successfully applied to overall amino acid composition-based subcellular localization prediction in the SubLoc program (Hua and Sun, 2001). Using the software LIBSVM (Lin, 2003), a similar

SVM was trained on 248 cytoplasmic sequences and 1054 non-cytoplasmic sequences. A query protein's amino acid composition is analyzed and used to assign the protein to one of the two categories: cytoplasmic or non-cytoplasmic.

3.2.6 Signal peptides

Signal peptides, short sequences present at the amino-terminus of many proteins, direct a protein for transport across the inner membrane (Bernstein, 1998). Thus the presence of a signal peptide implies that a protein is not resident in the cytoplasm. The bacterial SignalP training data, available at http://www.cbs.dtu.dk/ftp/signalp, were used to train a Hidden Markov Model (HMM) to identify signal peptide cleavage sites within the first 70 residues of a sequence. A probability value is assigned to the cleavage site and, if it exceeds a pre-assigned cutoff, a prediction of non-cytoplasmic is returned. If the p-value of the predicted cleavage site falls within a 'twilight zone', the signal peptide is then passed to an SVM trained on the same data, also capable of identifying signal peptides. If the SVM returns a result of signal peptide, a non-cytoplasmic prediction is returned. If no signal peptide is identified, the module returns an output of 'unknown', as the lack of a predicted signal peptide does not necessarily imply a cytoplasmic localization.

3.3 Evaluation of predictive module performance

All evaluations, with the exception of the Motifs module analysis, were carried out using 5-fold cross-validation. In *k*-fold cross-validation, the relevant dataset is partitioned randomly into *k* equally sized partitions and module development and evaluation is carried out *k* times, each time using one distinct partition as the testing set and the remaining *k*-1 partitions as the training set. Values of *k* of five and 10 are frequently employed in the evaluation of prediction methods, and choosing the smaller of these reduces the necessary computation. Performance evaluations are computed as the average of the total runs, thus the procedure prevents artificially inflated performance values.

After determining the numbers of true positive predictions (TP), false positive predictions (FP), and both true negatives (TN) and false negatives (FN), performance evaluation metrics can be calculated. Several such metrics exist (see section 5.2.1), however precision and recall were selected for the evaluation of PSORTb.

Precision is calculated as $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$. A precision value of 95% indicates that for every 100 predicted cytoplasmic proteins, five of these will be false positives, or non-cytoplasmic proteins. Recall, or sensitivity, is calculated as $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$ and reflects a method's ability to identify all true positive cases. A recall value of 95% indicates that for every 100 cytoplasmic proteins in the test set, five of these will be false negatives – in other words, they will be predicted as non-cytoplasmic when in fact they are cytoplasmic proteins.

SCL-BLAST was evaluated using the ePSORTdb v.1.0 dataset of 1443 proteins of known subcellular localization. PROSITE motifs were selected to yield a 100% precision value over the same dataset. The predictive power of HMMTOP was evaluated on the 268 integral inner membrane protein and the remaining 1175 non-inner membrane proteins in the dataset. Outer membrane protein motifs were evaluated using the 425 beta-barrel proteins in the dataset and the remaining 1018 non-outer membrane proteins. SubLocC was evaluated using the 248 cytoplasmic proteins in the dataset and the remaining 1054 non-cytoplasmic proteins—cytoplasmic proteins with a second, dual localization were not included in either class. The signal peptide module was trained using the SignalP dataset mentioned above and evaluated with the Menne et al. (2000) dataset of 426 signal peptides and 433 non-signal peptides. The precision and recall of each module are presented in Table 3.1. For modules capable of predicting multiple localization sites, the reported precision and recall values are averaged across the relevant localization sites.

Table 3.1: Predictive performance of PSORTb v.1.0 modules.

Module	Precision	Recall
SCL-BLAST	96.7	60.4
Motifs	100.0	6.5
HMMTOP	99.4	65.3
Outer Membrane Protein Motifs	100.0	23.6
SubLocC	78.6	74.2
Signal Peptides	87.0	98.2

Each module is implemented as a Perl script or as a Perl wrapper script that interfaces with another program. This modular design permits the simple introduction of additional analyses into the program. The program is developed under the GNU General Public Licence – an open source license – to encourage the open development and expansion of the tool.

3.4 Integration of the module output using a Bayesian network

The performance data for each module was used to construct a naive Bayesian network capable of generating a final probability value for each localization site given the output of each of the modules. A Bayesian network employs Bayes' theorem of conditional probabilities to calculate the likelihood of a particular scenario given that certain events have occurred. With respect to the localization prediction problem and PSORTb, the network allows us to calculate the probabilities of a protein being resident at each of the localization sites, given the output of specific modules. A score out of 10 is produced for each of the five possible localization sites, representing the calculated probability value multiplied by a factor of 10, with a high value reflecting high confidence that the given protein is resident in that subcellular location. The sites are ranked in descending order of probability. If none of the localization sites has a score >7.5, a prediction of 'unknown' is returned. This cutoff of 7.5 was determined empirically through a review of PSORTb's precision and recall at various cutoffs – 7.5 represented the point at which high (>95%) precision was obtained, while returning a significant number of results. Higher cutoffs reduced recall, while lower cutoffs reduced precision. A distribution of scores heavily favouring one site indicates the protein is likely to be resident there, while a distribution favouring two sites may indicate the protein has domains residing in more than one localization site. An even distribution of low scores is indicative of an unknown localization.

3.5 Performance of PSORTb v.1.0 vs. PSORT I

The overall performance of PSORTb was assessed using the ePSORTdb v.1.0 dataset comprising 1443 proteins of known localization. Precision and recall values were calculated per localization site and the overall precision and recall of the method was calculated based on the total number of true-positives (TP), false-positives (FP) and false-negatives (FN) over the five Gram-negative localization sites.

For the purposes of evaluation, predictions were considered to have been made if the PSORTb scoring system gave a score for a particular localization site of >7.5. Proteins resident at dual localization sites were considered to have been predicted correctly if one of their localization sites scored >7.5. For all evaluations precision was calculated as TP/(TP+FP) and recall was calculated as TP/(TP+FN). Results appear in Table 3.2. As shown in this table, this version of PSORTb was compared with the performance of PSORT I (Nakai and Kanehisa, 1991), again using the ePSORTdb v.10 dataset.

Table 3.2: Predictive performance of PSORTb v.1.0 compared to PSORT I.

	PSORT I		PSORTb v. 1.0	
Localization	**Precision (%)**	**Recall (%)**	**Precision (%)**	**Recall (%)**
C	59.7	75.4	97.6	69.4
CM	55.4	95.1	96.7	78.7
P	60.9	66.4	91.9	57.6
OM	65.3	54.5	98.8	90.3
EC	0.0	0.0	94.4	70.0
Total	59.6	60.9	96.5	74.8

PSORTb was designed to favour precision, with a focus on predicting results correctly rather than generating a prediction in every case. This is reflected in the precision and recall of both the modules and the overall program. With this emphasis on precision and implementation of an updated predictive strategy, the performance of PSORTb represents a significant improvement over the PSORT I program. Whereas a large increase in precision can be observed for each localization site, recall is reduced in certain cases, again reflective of the focus on returning a correct prediction rather than more predictions with lower confidence. This is especially evident for inner membrane proteins, for which a 16.4% decrease in recall is compensated for by a 41.3% increase in precision.

3.6 Release of PSORTb v.1.1

The beta version of the PSORTb software – PSORTb v.1.0 – was initially developed in January, 2003, however the tool was not released online until shortly before its publication in June, 2003. By this point, the ePSORTdb dataset had expanded to include an additional 131 Gram-negative bacterial proteins. This represented a significant update to the SCL-BLAST module; thus a new version number was assigned to the tool, such that the first publicly available PSORTb release was numbered 1.1.

PSORTb v.1.1 was made available at http://www.psort.org/psortb, a site that also contains links to many other tools, resources and articles of interest to the protein localization community. In this and all subsequent releases of the program, a user may submit one or more query proteins in FASTA format, either through a text box or through upload of a local file. A choice of three output formats is offered: Normal, consisting of a textual table showing the names of each module and their output, including details of the prediction, as well as the score distributions and a final prediction when possible; Tab-delimited (long), which contains the same information arranged in a spreadsheet-ready format, one line per query protein; and Tab-delimited (short), containing simply the protein identifier and the final prediction.

4 DEVELOPMENT AND RELEASE OF PSORTb V.2.0

Portions of this chapter have been previously published in the article "PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis", co-authored by J.L. Gardy, M.R. Laird, F. Chen, S. Rey, C.J. Walsh, M. Ester, and F.S.L. Brinkman in Bioinformatics, Volume 21, Issue 5 © 2005 Oxford University Press.

4.1 Summary

PSORTb v.1.1's predictive coverage and recall were low and the method was only applicable to Gram-negative bacteria. We set out to increase PSORTb's coverage while maintaining the existing precision level, and expand it to include Gram-positive bacteria. An expanded ePSORTdb database of proteins of known localization and new modules using frequent subsequence-based SVMs were introduced into PSORTb v.2.0. This version of the program attained a precision of 96% for Gram-positive and Gram-negative bacteria and displayed predictive coverage comparable to other tools for whole proteome analysis, representing a significant improvement over version 1.1.

4.2 Rationale

By analyzing features including: signal peptides, transmembrane helices, homology to proteins of known localization, amino acid composition and motifs, PSORTb v.1.0 and 1.1 attained a classification precision of 97%. However, the method did not extend to Gram-positive organisms and its predictive coverage when applied to whole proteomes – the number of proteins for which a prediction could be made – was low, with an average coverage of 28%. We therefore set out to expand PSORTb's predictive scope by introducing additional classification methods applicable to both Gram-positive and Gram-negative bacteria, and to increase the program's coverage while maintaining the existing standard of high precision.

4.3 Novel support vector machine modules

Support vector machine (Vapnik, 2000), is a kernel learning algorithm in which all data is mapped as vectors in *n*-dimensional feature space. Given training data from two classes (positive and negative), a SVM learns the optimal separating hyperplane which both separates the two classes and maximizes their distance from the hyperplane. In previous work on the applicability of SVMs to the localization classification problem, nucleotide or protein sequences have been modelled as vectors representing amino acid composition (Hua and Sun, 2001, Yu et al., 2004). We proposed, however, that the precision of an SVM could be improved by utilizing frequently-occurring subsequences rather than overall amino acid composition. Such common patterns within a group of proteins may indicate the site of a common biochemical mechanism or structural motif.

For each of the nine localization sites (five Gram-negative and four Gram-positive), a training dataset was created consisting of a positive and negative class. The positive class consisted of those proteins in ePSORTdb annotated as being resident at the localization site of interest, whereas the negative class consisted of the remainder of ePSORTdb.

4.3.1 Extraction of frequently occurring subsequences

Frequent subsequences were extracted from the protein sequences comprising each positive class. A subsequence, or pattern, is defined as frequent if it is found in at least a specified fraction – *MinSup*, or minimum support – of the proteins in ePSORTdb resident at a specific site. A frequent pattern has the form *X*X*…, in which each 'X' is a frequent subsequence made of consecutive amino acids, and each '*' is a VLDC (variable-length-don't-care) which may substitute for one or more letters when matching the pattern against a protein sequence. Subsequences capture the local similarity that may indicate of important structural or functional residues, while VLDCs compress the remaining irrelevant portions.

To find frequent subsequences, an efficient implementation of the generalized suffix tree (GST) (Wang et al., 1994) with some simple modifications was implemented. Suffix trees have been extensively used in string matching and are shown to be an effective data structure for finding common subsequences that run in linear time (Landau and Vishkin, 1989; Hui, 1992). Since each protein sequence is essentially a string of letters, generalized suffix trees can be easily applied to mine frequent subsequences among protein sequences. Each of the nine pattern extractions was performed over a range of *MinSup* values.

4.3.2 Development and implementation of SVM-based classifiers

SVMLight (Joachims, 2002) was used to implement nine SVMs whose feature spaces consisted of the frequent subsequences characteristic of a specific localization site. For each localization site, different SVMs were tested using different combinations of *MinSup* (range: 0.8%-13%) and kernel (linear, polynomial with degree = 2, radial basis function with γ=0.005). The *MinSup*/kernel combination giving the highest classification precision combined with a reasonable level of recall (> 40%) was selected for inclusion in PSORTb v.2.0. Variations in the margin error penalization parameter *C* were not evaluated, as our earlier collaborative work on the subject showed a negligible effect on precision and recall values (She et al., 2003). The final SVMs implemented in PSORTb v.2.0 utilize LibSVM (Lin, 2003). The cytoplasmic SVM replaces the SubLocC module of PSORTb v.1.0.

4.3.3 Evaluation of SVM-based classifiers

Table 4.1 summarizes the SVM classifiers selected for inclusion in PSORTb v.2.0.

Table 4.1: Parameters and performance of PSORTb v.2.0's SVM-based classifiers.

		SVM Parameters			Performance (%)	
Gram	**Module**	**MinSup (%)**	**Frequent patterns**	**Kernel**	**Precision**	**Recall**
Negative	CytoSVM-	0.5	39219	Linear	83.6	68.4
	CMSVM-	3	5645	Polynomial	96.9	69.6
	PPSVM-	1	27804	Polynomial	96.3	45.3
	OMSVM-	1	46688	Linear	94.6	85.3
	ECSVM-	2	35380	Polynomial	94.1	56.4
Positive	CytoSVM+	2	8214	Linear	86.5	79.9
	CMSVM+	2	250163	Linear	100.0	63.1
	CWSVM+	2	11610	Linear	95.7	55.6
	ECSVM+	5	23605	Polynomial	91.7	55.0

By using a feature space comprising frequent subsequences rather than amino acid composition, high precision classification across all localization sites was achieved. Although the precision values for the two cytoplasmic classifiers are the lowest of the nine values, the 84% precision achieved by the Gram-negative SVM represents a 5% increase relative to the cytoplasmic composition-based SVM SubLocC used in PSORTb v.1.1. The reduced precision associated with cytoplasmic proteins may be due to the extremely diverse nature of proteins found at this site – proteins found at other sites exhibit more functional and structural constraints, resulting in more unique and characteristic frequent subsequences. This is especially evident when classifying cytoplasmic membrane proteins – the frequent subsequences mined from this structurally and environmentally constrained group of proteins results in high precision classification.

We observed that as the *MinSup* value increased for each classifier, the number of frequent patterns decreased, as did precision; recall, however, remained comparatively stable. It was also noted that the best performance is not achieved at the smallest *MinSup* value – when the number of frequent subsequences exceeds a certain level, the performance of the SVM is degraded.

4.4 Further expansion of predictive capability

4.4.1 SCL-BLAST

PSORTb's SCL-BLAST module predicts the localization of a query sequence based on homology to a protein in the PSORTdb database of proteins of experimentally verified localization. It is therefore expected that a larger and more diverse database will lead to an increase in the program's recall. SCL-BLAST v.2.0 utilizes an updated version of the original PSORTdb database – Gram-positive queries are run against the subset of 576 new proteins of Gram-positive origin. Furthermore, we investigated whether subsets of the Gram-negative and Gram-positive database could be combined. For example, the cytoplasmic and cytoplasmic membrane sites were hypothesized to be functionally equivalent, such that a Gram-negative protein could be searched against a BLAST database containing both Gram-negative bacterial proteins and Gram-positive cytoplasmic and cytoplasmic membrane proteins. We examined whether a larger database with such combinations of proteins would increase recall even further.

We tested several combined databases using 5-fold cross-validation and found that higher recall and comparable precision was indeed achieved. For Gram-positive results, a database including Gram-negative cytoplasmic, cytoplasmic membrane and extracellular proteins yielded the best predictions. For Gram-negative queries, optimal results were achieved when the queries were searched against a database that included Gram-positive cytoplasmic and cytoplasmic membrane proteins – including extracellular proteins in the database resulted in several periplasmic proteins being falsely predicted as extracellular. Results of 5-fold cross-validation testing of SCL-BLAST v.2.0 for each localization site are shown in Table 4.2.

Table 4.2: Performance of the expanded SCL-BLAST module in PSORTb v.2.0.

		Performance (%)	
Gram	**Localization**	**Precision**	**Recall**
Negative	C	88.8	39.9
	CM	97.4	62.0
	P	94.4	68.8
	OM	99.4	90.5
	EC	97.3	77.4
	Total	96.4	68.6
Positive	C	96.6	58.8
	CM	96.8	59.8
	CW	91.9	56.7
	EC	95.5	57.7
	Total	95.7	58.4

The Gram-negative version of the module retains the 96% precision exhibited in PSORTb v.1.1, and improves the recall by 8%. The new Gram-positive version also displays precision of 96%, and recall of 58%, with the lower recall likely due to the smaller Gram-positive bacterial protein dataset. It is important to note, however, that such recall values are not to be expected when SCL-BLAST is applied to datasets containing a large number of hypothetical proteins, due to their lack of similarity to proteins in the SCL-BLAST database.

We also introduced an exact match filter to detect if a user's query protein is already in the database – if a query protein displays 100% identity to a protein in PSORTdb with a difference between query and subject length of no more than one character (to account for some users' removal of the initial "f-Methionine" residue), the SCL-BLASTe subroutine returns the localization site associated with the subject protein. In cases in which an exact match is identified, the query protein is not analyzed by subsequent modules, enabling a result to be returned faster.

4.4.2 Motifs and profiles

In PSORTb v.1.1, the Motif module scanned a query sequence for the presence of any one of 26 PROSITE motifs indicative of specific Gram-negative localization sites. In PSORTb v.2.0, the module was expanded to include 44 Gram-negative motifs derived from PROSITE v.18 (Hulo *et al.*, 2004), covering all but the cytoplasmic localization site, and 25 Gram-positive motifs covering all 4 localization sites. The complete list of motifs is provided in Appendix A. Each motif was checked against ePSORTdb to ensure that it produces no false positive results. Two motifs used in PSORTb v.1.1 were removed from v.2.0 due to the occurrence of false positives when examined against the expanded ePSORTdb dataset.

PSORTb v.2.0 also includes a Profile module, in which localization-specific profiles derived from PROSITE v.18 were selected to generate 100.0% precise predictions against ePSORTdb. Each profile is similar to a motif but with position-specific weighting information included, such that more degenerate sequences can be retrieved than via the strict pattern-matching of the Motif module. Six profiles were selected, four of which identify both Gram-negative and Gram-positive cytoplasmic and cytoplasmic membrane proteins, and two of which are specific to the Gram-positive cell wall and extracellular sites. Profiles are provided in Appendix C.

4.4.3 Signal peptides

A separate signal peptide prediction module was trained using Gram-positive data derived from the same source as the original Gram-negative training data, at http://www.cbs.dtu.dk/ftp/signalp.

4.5 Updates to module output integration

As in version 1.1, the modules' predictions are weighted and integrated using a Bayesian network in order to generate the final prediction, which comes in the form of a score distribution. When a single localization site displays a score of 7.5 or greater, that site is returned as a final prediction. New to version 2.0 is multiple localization flagging – if two sites return high scores, a flag of "This protein may have multiple localization sites" is appended to the final prediction. This flag is triggered when a site scores between 4.0 and 7.49 for Gram-negative bacterial proteins, and between 5.0 and 7.49 for Gram-positive bacterial proteins. If no site scores above 4.0 or 5.0, depending on the class, a localization site of "Unknown" is returned. Score cutoffs were determined again by evaluating which cutoff yielded the highest precision classification while returning a reasonable number of results, although this time the dataset used in evaluation comprised proteins with multiple localizations. PSORTb's emphasis is on precision, and returning a result of "Unknown" when not enough information is available to make a prediction avoids potential false positive results. Figure 4.1 illustrates the final architecture of PSORTb v.2.0, and Figure 4.2 provides examples of both normal and tab-delimited terse format output of the program.

Figure 4.1: Organization of PSORTb v.2.0.

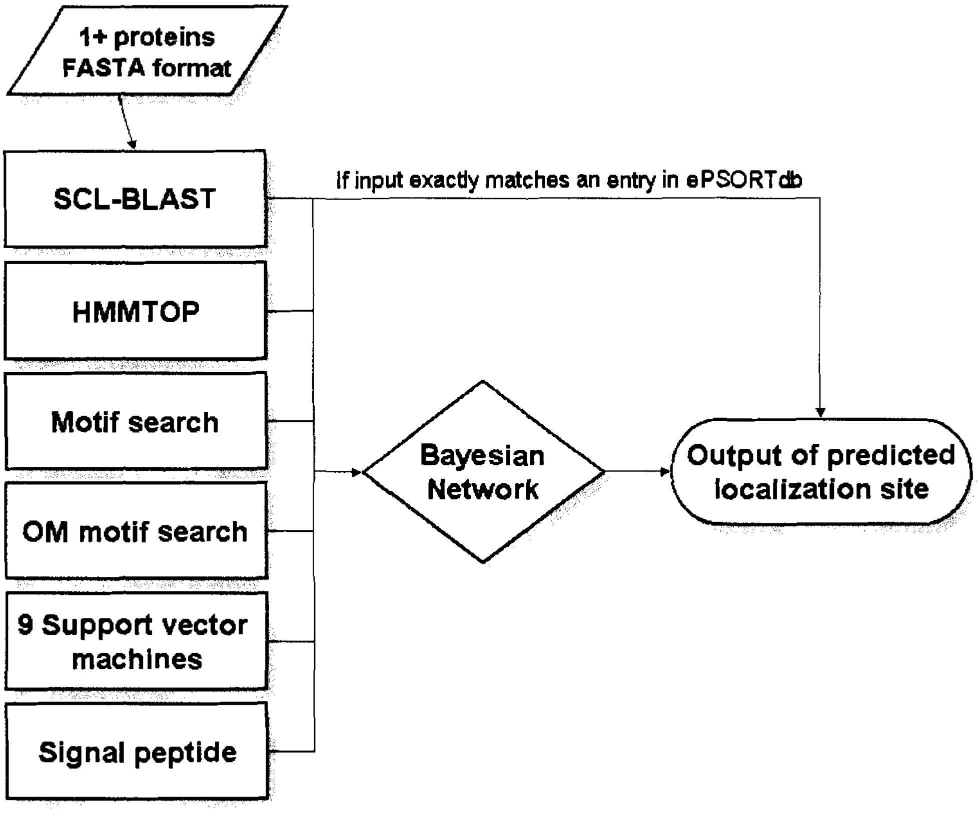
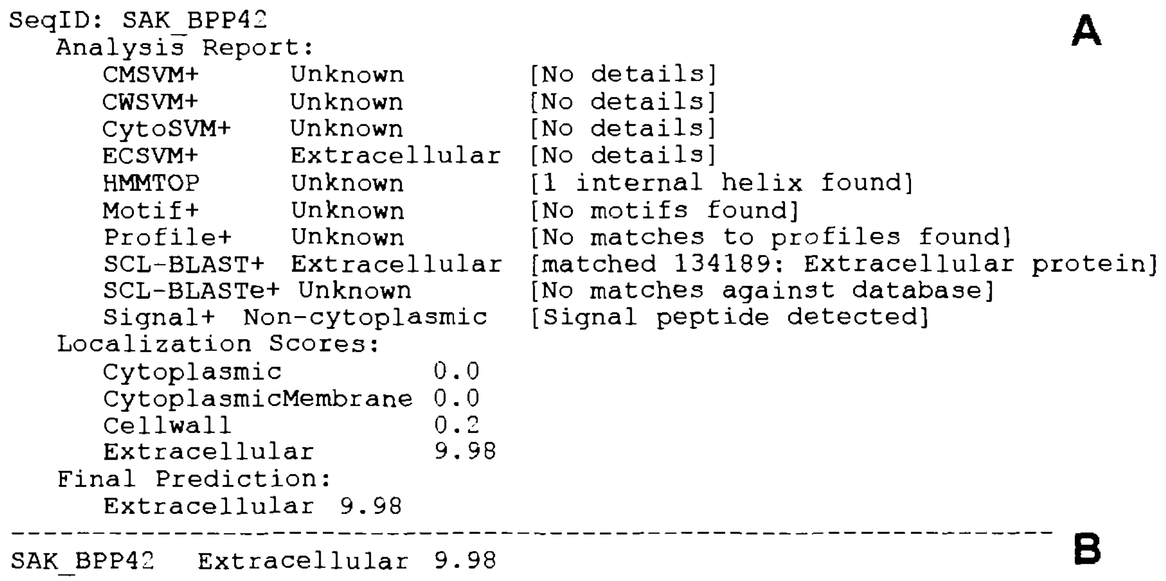

1+ proteins
FASTA format
SCL-BLAST
HMMTOP
Motif search
OM motif search
9 Support vector machines
Signal peptide
If input exactly matches an entry in ePSORTdb
Bayesian Network
Output of predicted localization site

Figure 4.2: Normal (A) and tab-delimited short (B) format PSORTb v.2.0 output.



```
SeqID: SAK_BPP42                                                    A
   Analysis Report:
      CMSVM+       Unknown        [No details]
      CWSVM+       Unknown        [No details]
      CytoSVM+     Unknown        [No details]
      ECSVM+       Extracellular  [No details]
      HMMTOP       Unknown        [1 internal helix found]
      Motif+       Unknown        [No motifs found]
      Profile+     Unknown        [No matches to profiles found]
      SCL-BLAST+   Extracellular  [matched 134189: Extracellular protein]
      SCL-BLASTe+ Unknown         [No matches against database]
      Signal+ Non-cytoplasmic     [Signal peptide detected]
   Localization Scores:
      Cytoplasmic          0.0
      CytoplasmicMembrane  0.0
      Cellwall             0.2
      Extracellular        9.98
   Final Prediction:
      Extracellular  9.98
-------------------------------------------------------------------
SAK_BPP42   Extracellular 9.98                                      B
```

4.6 Performance of PSORTb v.2.0

4.6.1 Precision and recall

Five-fold cross-validation was used to evaluate the Gram-negative and Gram-positive versions of PSORTb v.2.0. Precision and recall values for each localization site were calculated for both proteins annotated as having a single localization site (Table 4.3) and dual localization sites (Table 4.4).

Table 4.3: PSORTb v.2.0 performance on singly-localized proteins.

		Performance				
Gram	**Localization**	**TP**	**FP**	**FN**	**Precision (%)**	**Recall (%)**
Negative	C	195	15	83	92.9	70.1
	CM	286	14	23	95.3	92.6
	P	191	9	85	95.5	69.2
	OM	371	10	20	97.4	94.9
	EC	150	4	40	97.4	78.9
	Total	1193	52	251	95.8	82.6
Positive	C	168	5	26	97.1	86.6
	CM	94	3	9	96.9	91.3
	CW	54	3	7	94.7	88.5
	EC	124	8	59	93.9	67.8
	Total	440	19	101	95.9	81.3

Table 4.4: PSORTb v.2.0 performance on multiply-localized proteins.

		Performance				
Gram	**Localization**	**TP**	**FP**	**FN**	**Precision (%)**	**Recall (%)**
Negative	C/CM	11	2	5	84.6	68.8
	CM/P	34	1	17	97.1	66.7
	P/OM	2	2	0	50.0	100.0
	OM/EC	76	1	2	98.7	97.4
	Total	123	6	24	95.3	83.7
Positive	C/CM	12	6	3	66.7	80.0
	CM/CW	6	0	14	100.0	30.0
	Total	18	6	17	75.0	51.4

For a protein resident at X and Y localization sites, a true positive (TP) is a prediction of either X, Y, or X/Y. A false positive (FP) is all multiply-localized proteins not resident at X or Y which are predicted as X, Y, or X/Y. A false negative (FN) is all X/Y proteins not predicted as neither X, Y, nor X/Y.

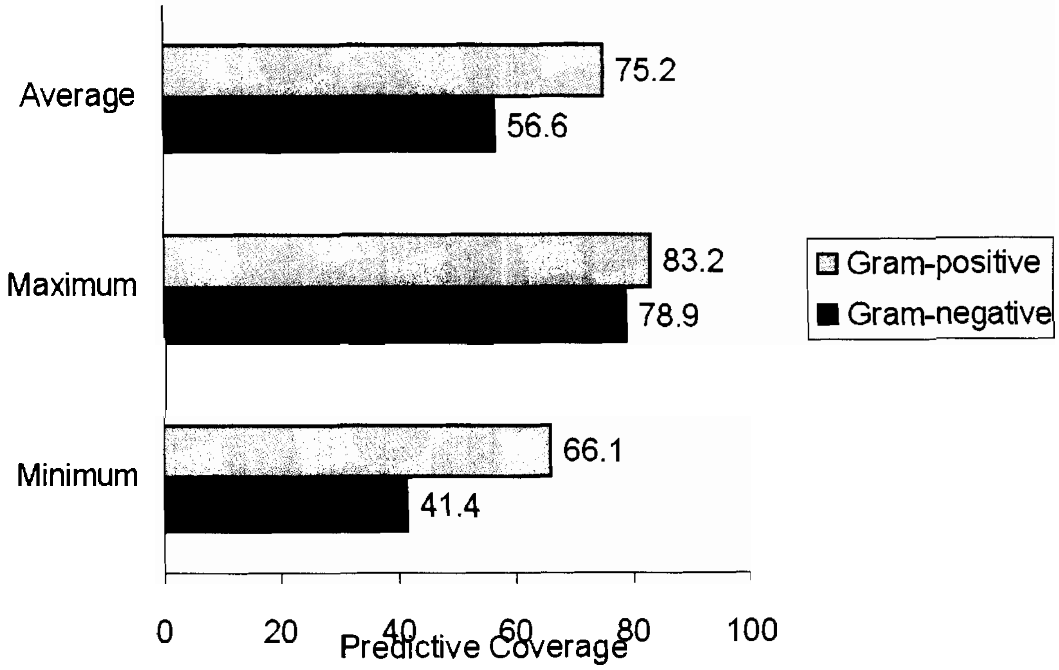
On single localization proteins, PSORTb v.2.0 attained precision values of 96% for both classes of organisms, and recall of 83% and 81% for Gram-negative and Gram-positive bacterial proteins, respectively. It was observed that precision values remained relatively constant across localization sites, while the recall was highest for membrane proteins, likely due to their conserved structural motifs readily identifiable by the frequent subsequence-based SVMs, HMMTOP and OMPMotif modules. The Gram-negative version of PSORTb v.2.0 exhibits a 0.7% drop in precision relative to PSORTb v.1.1, however an 8% increase in recall is observed.

Performance of the program on proteins annotated as having dual localization sites is comparable to the performance for singly localized proteins with respect to Gram-negative organisms, with a precision of 95% and recall of 84%. However, it was noted that the overall precision for Gram-positive multiply localized proteins was only 75%. Upon inspection, it became apparent that this was due to six annotated cytoplasmic membrane/cell wall proteins being predicted as cytoplasmic. Noting that singly-localized cytoplasmic membrane and cell wall proteins were infrequently mispredicted as cytoplasmic, these six proteins were investigated. While experimental evidence supporting a possible additional localization of cytoplasmic was only found for one of the six proteins – *B. subtilis* ComGG (Chung *et al.*, 1998) – the other five proteins include enzymes and heat shock proteins, for which cytoplasmic or peripheral membrane associated localizations are not uncommon. It may be that rather than making mispredictions, PSORTb is detecting a more complex pattern of localization for certain proteins.

4.6.2 Predictive coverage and storage of genome-wide predictions

The predictive coverage of a method refers to the number of proteins in a given proteome for which the method returns a prediction. The measured recall of a program when evaluated using 5-fold cross validation does not give an accurate reflection of the predictive coverage because the training and testing data consists of a number of well-characterized proteins, thus a large number of predictions are possible. However, proteomes contain a large number of hypothetical proteins, which often do not contain enough information for a prediction to be generated. We therefore set out to measure PSORTb v.2.0's performance when applied to whole proteomes, with the expectation that an increase in the 28% average coverage of version 1.1 would be observed. Figure 4.3 summarizes PSORTb v.2.0's predictive coverage when applied to the analysis of 162 Gram-negative genomes and 74 Gram-positive genomes. The average coverage across each class of proteomes is shown, as is the maximum and minimum coverage values obtained for a single proteome in each class.

Figure 4.3: Predictive coverage of PSORTb v.2.0 when applied to complete genomes.



The Gram-positive version of the program displays higher predictive coverage than the Gram-negative version due to the higher recall associated with the Gram-positive cytoplasmic SVM. Cytoplasmic proteins represent the largest class of proteins within the cell, and an improved ability to identify these proteins results in a higher overall coverage.

PSORTb v.2.0 was released on June 17, 2004 at http://www.psort.org/psortb. On August 12, a second component of the PSORTdb database was also released – cPSORTdb. cPSORTdb stores PSORTb v.2.0 predicted localization sites for all sequence bacterial genomes available through NCBI. As of January, 2006, cPSORTdb contains 689,275 proteins representing 236 organisms.

Most fields present in an ePSORTdb record are also present in a cPSORTdb record, however the latter also contains fields for NCBI's genome accession number, strain designation, and cross-linking if the record in question is also present in ePSORTdb. By definition, cPSORTdb also must contain the PSORTb predictive output, which is stored in a series of unique fields. Like ePSORTdb, cPSORTdb can be queried with text, browsed, or searched using BLAST, and the results are returned in a similar format.

5 EVALUATING PSORTb v.2.0'S PRECISION IN THE CONTEXT OF RECENT COMPUTATIONAL PREDICTIVE METHODS

Portions of this chapter have been previously published in the article "PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis", co-authored by J.L. Gardy, M.R. Laird, F. Chen, S. Rey, C.J. Walsh, M. Ester, and F.S.L. Brinkman in Bioinformatics, Volume 21, Issue 5 © 2005 Oxford University Press.

5.1 Summary

Beginning with the release of SubLoc, a number of research groups turned their attention to the development of localization prediction tools. An objective comparison of the performance of these publicly available was carried out using a set of proteins not included in the training data of any of the methods under review. It was thus shown that PSORTb represents the highest precision bacterial localization prediction method released to date.

5.2 Alternative localization prediction tools

The evaluation of PSORT I described in the PSORTb v.1.0 manuscript (Gardy et al., 2003) dramatically illustrated the need for an improved localization prediction tool. Consequently, a number of research groups undertook development of novel predictive algorithms. As of January, 2006 seven bacterial protein localization prediction tools, including PSORTb v.2.0, have been made available. While these seven methods vary with respect to the algorithms employed, the number of localization sites they are able to assign a protein to, predictive performance and user interface, they all offer an improvement in precision over PSORT I.

5.2.1 A note regarding performance metrics

Before examining each of the methods, an explanation of the various performance metrics used by the methods' authors is required. The choice of metrics depends primarily on the background of the authors and the aspect of their method which they wish to emphasize. All metrics, however, rely on four basic statistics – true positives (TP), false negatives (FN), false positives (FP) and true negatives (TN).

Predictive methods developed by biologists tend to emphasize the importance of "quality", or correct predictions, over "quantity", or a high number of predictions. A barometer of a method's ability to generate correct predictions is the precision metric, calculated as $\frac{\mathrm{TP}}{\mathrm{TP+FP}}$. Precision values are typically reported together with a method's recall. Recall, also referred to as sensitivity, is calculated as $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$ and reflects a method's ability to identify all true positive cases.

A small number of papers from the computer science domain report only a single metric – accuracy. To many, the use of the word accuracy implies a measure of quality, or $\frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}$. However, many computer scientists use a different definition of accuracy. This version of accuracy is the same as our earlier definition of recall, or $\frac{\mathrm{TP}}{\mathrm{TP+FN}}$, and rewards methods that generate a large quantity of predictions, thus should not be used as an estimate of a method's false positive rate.

All of the predictive methods described below also report some sort of confidence measure along with their final prediction, typically under the name "confidence", "quality" or "reliability index". These measures, which range from scores on a five- or ten-point scale to percentages, should only be used to compare the confidence level of multiple predictions from a single server. Because the calculation of these measures varies widely between servers, it is important to remember they are relative measures and thus it is not possible to compare these quality scores between servers.

5.2.2 Proteome Analyst

Proteome Analyst (Lu et al., 2004) was trained and tested using Gram-negative bacterial sequences extracted from the first release of ePSORTdb as well as SwissProt entries with annotated localization information. The complete Gram-negative dataset consists of 3174 sequences, and the Gram-positive dataset 1541 sequences. The authors report precision of 95.9% and 94.6%, and recall of 93.4% and 91.4% for the Gram-negative and Gram-positive classifiers, respectively. Coverage analysis was performed using one genome from each class of bacteria, with Proteome Analyst achieving 75.6% coverage for the Gram-negative bacterium *Pseudomonas aeruginosa* and 67.2% coverage for the Gram-positive bacterium *Bacillus subtilis*. When PSORTb v.2.0 was used to analyze the same organisms, it attained coverage of 68.1% for *P. aeruginosa* and 76.5% for *B. subtilis*.

An analysis based on these two proteomes suggests that while Proteome Analyst attains higher coverage on a Gram-negative organism, PSORTb v.2.0 generates more predictions for a Gram-positive proteome. However, because of the small sample size and the fact that these two organisms are quite well-annotated, the true coverage of each method cannot accurately be compared.

Proteome Analyst's markedly different approach and high recall make it an excellent complement to PSORTb. Like PSORTb, Proteome Analyst does not force predictions, resulting in high precision classification. A caveat to the method, however, is that in order to generate a prediction, a query protein must have homologs in the SwissProt database – hence not every protein encoded in a genome will return a result. Furthermore, predictions for proteins that are similar to known, well-studied proteins will also be of higher confidence than those for proteins whose homologs are not well-annotated.

5.2.3 Amino acid composition support vector machine-based methods

Composition-based SVMs are capable of making predictions when no prior information – homology or the presence of a particular sequence feature, for example – about a protein is available. However, because predictions are returned for all submitted queries – even if the result is not a highly confident one – the precision of these methods is significantly lower than that of PSORTb and Proteome Analyst.

SubLoc (Hua and Sun, 2001) was trained on the Reinhardt and Hubbard dataset derived from SwissProt, comprising 997 bacterial sequences (Reinhardt and Hubbard, 1998), and the authors report an accuracy value of 91.4% when the method was evaluated using this dataset.

CELLO (Yu et al., 2004) was trained and tested using the 1443 Gram-negative bacterial proteins in the first release of ePSORTdb (Gardy et al., 2003). The authors report 88.9% accuracy, with predictions generated for all queries. The method was recently updated to permit the analysis of Gram-positive sequences, however information on the training data used and the performance of the method is unavailable.

LOCtree (Nair and Rost, 2005) was trained and tested using a dataset created by the program's authors comprising 672 bacterial proteins. The authors report accuracy of 83% for Gram-negative bacteria and 90% for Gram-positive bacteria. However, because LOCtree does not discriminate between Gram-negative and Gram-positive organisms, the potential exists for a Gram-positive protein to be mistakenly classified as periplasmic.

P-CLASSIFER (Wang et al., 2005) was trained and tested on a dataset derived from the first release of ePSORTdb comprising 1441 proteins, achieving a recall of 89.8%, and PSLpred (Bhasin et al., 2005) was trained using 1443 proteins from the first release of ePSORTdb, achieving 91.2% accuracy.

5.3 Comparison of precision and recall

Because of the disparity in the performance metrics reported by the authors of each of the above methods, it is difficult to accurately assess the performance of each program using information provided in the original manuscripts. Instead, an independent comparison of the tools using a set of proteins not contained in the training data of any of the methods is required.

We compared the performance of PSORTb, Proteome Analyst, CELLO, PSLpred and P-CLASSIFIER using 144 novel Gram-negative bacterial proteins not contained in the training data of any of the methods. SubLoc and LOCtree were not evaluated, as these methods only predict three localization sites rather than five and were trained using data which may overlap with the testing data used herein. Results are summarized in Table 5.1.

Table 5.1: An independent comparison of the performance of five bacterial subcellular localization prediction methods using a test set of 144 novel Gram-negative bacterial proteins.

Localization	Program	TP	FP	FN	Precision (%)	Recall (%)
C	PSORTb	22	1	8	95.7	73.3
	CELLO	27	9	3	75.0	90.0
	Proteome Analyst	22	1	8	95.7	73.3
	P-CLASSIFIER	27	7	3	79.4	90.0
	PSLpred	28	3	2	90.3	93.3
CM	PSORTb	39	1	3	97.5	92.9
	CELLO	35	4	7	89.7	83.3
	Proteome Analyst	40	6	2	87.0	95.2
	P-CLASSIFIER	38	6	4	86.4	90.5
	PSLpred	41	2	1	95.3	97.6
P	PSORTb	26	0	6	100.0	81.3
	CELLO	16	6	16	72.7	50.0
	Proteome Analyst	29	1	3	96.7	90.6
	P-CLASSIFIER	13	3	19	81.3	40.6
	PSLpred	20	4	12	83.3	62.5
OM	PSORTb	34	1	5	97.1	87.2
	CELLO	24	3	15	88.9	61.5
	Proteome Analyst	34	0	5	100.0	87.2
	P-CLASSIFIER	24	3	15	88.9	61.5
	PSLpred	23	2	16	92.0	59.0
EC	PSORTb	1	0	0	100.0	100.0
	CELLO	1	19	0	5.0	100.0
	Proteome Analyst	1	6	0	14.3	100.0
	P-CLASSIFIER	1	22	0	4.3	100.0
	PSLpred	1	20	0	4.8	100.0
Total	PSORTb	122	3	22	97.6	84.7
	CELLO	103	41	41	71.5	71.5
	Proteome Analyst	126	14	18	90.0	87.5
	P-CLASSIFIER	103	41	41	71.5	71.5
	PSLpred	113	31	31	78.5	78.5

5.4 Conclusions

Our assessment of the available bacterial protein localization prediction tools reveals that PSORTb v.2.0 achieves the highest precision, while Proteome Analyst achieves the highest recall, together with excellent precision. Both programs outperform the tools that force predictions, however of these SVM-based methods, PSLpred appears to yield the highest precision and recall.

Our results illustrate that multi-component predictive methods and those that take into account biological knowledge (in the form of SWISS-PROT annotations) generate higher quality predictions than do simple SVM-based methods.

6 ANALYSIS I: *PSEUDOMONAS AERUGINOSA* EXPORTED PROTEINS

Portions of this chapter have been previously published in the article "Genome-wide identification of Pseudomonas aeruginosa exported proteins using a consensus computational strategy combined with a laboratory-based PhoA fusion screen", co-authored by S. Lewenza, J.L. Gardy, F.S.L. Brinkman, and R.E.W. Hancock in Genome Research, Volume 15, Issue 2 © 2005 Cold Spring Harbor Laboratory Press.

6.1 Summary

The Gram-negative pathogen *Pseudomonas aeruginosa* encodes multiple protein export systems, the substrates of which contain export signals such as N-terminal signal peptides. Here we describe the first genome-wide computational and laboratory screen for N-terminal signal peptides in this important opportunistic pathogen. The computational identification of signal peptides was based on a consensus between multiple predictive tools and showed that 38% of the P. aeruginosa PAO1 proteome was predicted to encode exported proteins, most of which utilize cleavable type I signal peptides or uncleavable transmembrane helices. In addition, known and novel lipoproteins (type II), twin arginine transporter (TAT), and prepilin peptidase substrates (type IV) were also identified. A laboratory-based screen using the alkaline phosphatase (PhoA) fusion method was then used to test our predictions. In total, 310 nonredundant PhoA fusions were successfully identified, 296 of which possess a predicted export signal. Analysis of the PhoA fusion proteins lacking an export signal revealed that three proteins have alternate translation start sites that encode signal peptides, two proteins may use an unknown export signal, and the

remaining nine proteins are likely cytoplasmic proteins and represent false positives associated with the PhoA screen. Our approach to identify exported proteins illustrates how computational and laboratory-based methods are complementary, where computational analyses provide a large number of accurate predictions while laboratory methods both confirm predictions and reveal unique cases meriting further analysis.

6.2 Rationale

The release of the high-precision, high-coverage PSORTb v.2.0 tool permitted us to couple a comprehensive computational analysis of localization in a genome to a similarly large-scale laboratory-based analysis. It was decided to analyze exported proteins in *Pseudomonas aeruginosa* using both PSORTb and other high quality predictive methods and compare our results to laboratory-derived localization information.

The completion of the *Pseudomonas aeruginosa* PAO1 genome sequence has provided many insights into the biology and pathogenesis of this organism and serves as the starting point for genome-wide studies of this important opportunistic pathogen (Stover et al., 2000). *P. aeruginosa* is the primary cause of chronic lung infections and mortality in patients with cystic fibrosis and is also the third most frequently isolated nosocomial pathogen, causing approximately 10% of all hospital-acquired infections (Fridkin and Gaynes, 1999; Govan and Deretic, 1996; Hancock and Speert, 2000). *Pseudomonas* infections are difficult to treat due to the high intrinsic antibiotic resistance of this organism, which is attributed to low outer membrane permeability coupled with additional resistance mechanisms, including active drug efflux and antibiotic modification (Hancock and Speert, 2000).

A major subset of the *P. aeruginosa* proteome is dedicated to proteins that are exported out of the cytoplasm to the cell envelope – the cytoplasmic membrane, the periplasm and the outer membrane – or that are secreted out of the cell to the extracellular environment (Guina et al., 2003; Nouwens et al., 2000). This subset of proteins is involved in essential cellular processes that include cell wall assembly, nutrient uptake, virulence, antibiotic resistance, pili and flagella biogenesis, immunogenicity, adherence, energy generation and environmental sensing. The importance of these proteins is illustrated by the fact that many cell envelope proteins are the targets of current antimicrobials (Drew et al., 2003) or vaccines and thus identifying novel envelope proteins may provide new targets for drug discovery and immunoprophylaxis (Cachia and Hodges, 2003).

P. aeruginosa proteins destined to non-cytoplasmic subcellular localizations utilize various protein export systems, as recently reviewed in Ma et al. (2003). The Sec machinery facilitates the majority of protein transport across the cytoplasmic membrane (Filloux et al., 1998; Pugsley, 1993a). Proteins may be recognized by the SecB chaperone after translation, maintaining their appropriate conformation to permit recognition by the SecYEG translocation machinery (de Gier and Luirink, 2001; Drew et al., 2003; Pugsley, 1993a), or they may be recognized by the signal recognition particle (SRP), and directed ultimately to the SecYEG translocase (Bernstein, 2000; de Gier and Luirink, 2001; Drew et al., 2003). Export systems that are Sec-independent are also found in *P. aeruginosa*, including the twin-arginine translocation (TAT) pathway, which is responsible for the translocation of certain pre-folded proteins (Ochsner et al., 2002; Voulhoux et al., 2001), as well as type I and type III secretion systems that translocate proteins across the cytoplasmic and outer membranes in a single step (Ma et al., 2003).

The availability of the *P. aeruginosa* PAO1 genome sequence, combined with knowledge of the defined structures and motifs found in most N-terminal signal peptides, permitted us to perform a genome-wide computational survey of proteins that use N-terminal signal peptides for export out of the cytoplasm. Our definition of "exported protein" includes all proteins exported out of the cytoplasm, including those incorporated into the cytoplasmic membrane through the presence of transmembrane helices. Laboratory-based surveys of signal peptide-encoding genes was also possible through the use of the alkaline phosphatase (PhoA) fusion technique (Manoil and Beckwith, 1985). In this approach, signal peptide-containing genes are fused in frame with a truncated *`phoA* gene lacking its native signal peptide. The signal peptide in the fused genomic fragment targets the PhoA moiety across the inner membrane to the periplasm, where alkaline phosphatase folds and becomes enzymatically active.

We performed a combined computational and laboratory survey of *P. aeruginosa* signal peptide-encoding genes, first screening the *P. aeruginosa* PAO1 genome for potential export candidates using computational techniques, and then implementing a random cloning PhoA fusion screen to test our predictions. The analysis represents the most comprehensive analysis of exported proteins in *P. aeruginosa* to date, and clearly illustrates the utility of this combined approach to genome-scale studies.

6.3 Computational prediction of the exported fraction of the *P. aeruginosa* proteome

6.3.1 Methods

The version of the *Pseudomonas aeruginosa* PAO1 genome used in the present analysis was downloaded from http://www.pseudomonas.com, updated June 10 2004. This version of the genome annotation contains 5570 predicted proteins.

Type I signal peptides were predicted by a consensus approach utilizing four signal peptide prediction methods: SignalP v.3.0's neural network and hidden Markov model implementations (Bendtsen et al., 2004), LipoP v.1.0 (Juncker et al., 2003), and Phobius (Kall et al., 2004). A protein was noted as having a type I signal peptide if three or more methods predicted one, and as not having a signal peptide if three or more methods did not predict one. Fifty-seven proteins were noted as having possible type I signal peptides, representing cases where two methods predicted a signal peptide while two methods did not. Type II signal peptides were predicted exclusively by the program LipoP.

Sequences of type IV prepilin precursors and related proteins (Lory, 1994) were used to construct the motif G[FIMLSY][TS][LT][ILVP]E. The motif was then used to scan the *P. aeruginosa* PA01 genome for possible prepilin peptidase substrates. Downstream hydrophobic tracts were identified using the Kyte and Doolittle hydrophobicity scale (Kyte and Doolittle, 1982).

Possible TAT substrates were identified by searching for occurrences of the RRXFL[KR] motif (Chaddock et al., 1995), where a protein exhibited the dual arginines as well as matches to at least two of the FL[KR] residues. This set of proteins was filtered to remove proteins with little to no hydrophobic character in the region immediately C-terminal to the TAT motif, again using the Kyte-Doolittle hydrophobicity index.

Proteins utilizing a transmembrane helix for targeting were identified by both Phobius and TMHMM (Krogh et al., 2001). In the absence of a strongly predicted signal peptide (three or more predictions) and given a prediction of two or more transmembrane helices by either Phobius or TMHMM, a protein was annotated as using a TMH for export.

Predictions for all classes of export signal are summarized in Table 6.1 (see also Appendix D).

Table 6.1: Occurrence of computationally predicted export signals in *P. aeruginosa*.

Type of export signal	Number	% of genome	% of export signals
Type I	801	14.4	37.7
Possible type I	57	1.0	2.7
Type II (lipoprotein)	185	3.3	8.7
Type IV(prepilin)	23	0.4	1.1
TAT	15	0.3	0.7
Transmembrane helix	1042	18.7	49.1
Total with export signals	2123	38.1	-
No export signal	3447	61.9	-

6.3.2 Results: Type I signal peptides

The majority of proteins using an N-terminal signal peptide for export are substrates of the Sec pathway and are cleaved by signal peptidase I. In the *P. aeruginosa* genome, 801 (14.4%) proteins were predicted by at least three of the four predictive methods to contain a cleavable type I signal peptide. The programs typically agree in their predictions, with 518 out of 801 signal peptides having four identically predicted cleavage sites and an additional 56 signal peptides with three identically predicted cleavage sites.

In addition to the 801 proteins with strongly predicted signal peptides, 57 proteins yielded inconclusive results with only two out of four methods making a signal peptide prediction. These proteins were therefore classified as possible type I signal peptides. This list includes the known TAT substrate phospholipase PlcH (PA0844) (Voulhoux et al., 2001), a lectin protein thought to be anchored in the outer membrane (PA3361) and the probable outer membrane protein OprC (PA3790). Interestingly, this list also contains three regulatory proteins which are predicted to be cytoplasmic (PA1949, PA1998, PA2267) by PSORTb v.2.0 (Gardy et al., 2005). These three proteins likely represent the small number of false positives inherent in any predictive technique.

6.3.3 Results: Type II signal peptides

We identified 185 proteins, or 3.3% of the genome, as potentially containing a type II signal peptide. This number is considerably higher than the 76 PSORT I-predicted lipoproteins (Nakai and Horton, 1999) annotated at http://www.pseudomonas.com, likely due to the improved identification algorithm of LipoP. LipoP also reports the +2 residue of each predicted lipoprotein, as this residue is thought to act in targeting of the lipoprotein to one or the other membrane. The majority of *P. aeruginosa* lipoproteins contain Ser (54), Ala (49), or Gly (32) residues at this position, indicating that they are likely localized to the outer membrane (Yamaguchi et al., 1988).

6.3.4 Results: Type IV signal peptides

By scanning for the occurrence of a GFTLIE-like motif preceding a stretch of hydrophobic residues, 13 candidate type IV signal peptides were initially identified, all of which were present in proteins annotated as pseudopilins, type II secretion proteins, general secretion pathway proteins, pilins, or fimbrial subunits – classes of proteins known or suspected to be processed by prepilin peptidase.

The 13 predicted prepilin peptidase substrates occurred in clusters along the genome. Reasoning that neighbouring proteins might exhibit type IV-like signal sequences missed in the initial scan, the 10 proteins both upstream and downstream of the clusters were manually inspected. A further 10 sequences representing possible prepilin peptidase substrates were identified in this fashion. In total, 23 proteins were predicted to represent prepilin peptidase substrates, six of which – PA2671, PA2672, PA2673, PA2674, PA2675 and PA4554 – have not been previously described.

6.3.5 Results: TAT signal peptides

Proteins exported via the TAT machinery display an RRXFL[KR]-like motif at their N-terminus, which otherwise contains a leader sequence that resembles a tripartite signal peptide (Berks, 1996). The *P. aeruginosa* genome has been scanned for the presence of TAT substrates in two previous studies. Ochsner et al. (2002) identified 18 putative substrates through a manual inspection approach. Their criteria included: the presence of twin arginines, a match to at least one of the remaining residues in the motif, a hydrophobic tract following the motif, and an AXA cleavage site. Dilks et al. (2003) implemented their TATFIND v.1.2 program to identify 57 putative substrates, using the criteria of the presence of an XRRXXX motif within residues 1-34 as well as an uncharged region of 13 or more residues downstream of the twin arginines.

In the present analysis, proteins with predicted type I or II signal peptides were scanned for the presence of the TAT motif RRXFL[KR] immediately N-terminal to a stretch of hydrophobic residues. This identified 14 proteins with type I signal peptides that also possess TAT motifs, while one protein with a predicted type II signal peptide contains the TAT motif (PA4712). However, several putative TAT substrates reported in the earlier two studies were not identified in this analysis. This is attributed to the fact that many of these proteins do not contain traditional type I or type II signal peptides. Thus a second analysis was performed in which we eliminated the requirement for a predicted signal peptide. When the entire genome was searched without the signal peptide filtering step, an additional 12 putative TAT substrates were found.

In total, 27 potential TAT substrates were identified, 10 of which were not described in either of the two previous studies. The 10 novel substrates predicted include: FepD, HcnC, Sss, GlcE and 6 hypothetical or conserved hypothetical proteins. Sss, however, is annotated as a site-specific recombinase and likely represents a false positive associated with the scanning procedure. These proteins may have been missed in the previous two analyses due to the requirement for an AXA cleavage site in the Ochsner et al. (2002) study, and the minimum of 13 uncharged residues downstream of the twin arginines required by TATFIND.

Eight proteins identified by Ochnser et al. (2002) were not found in the present study. Inspection of these revealed that the proteins either exhibit weakly hydrophobic regions downstream of their motifs or have less than two residues in common with the FL[KR] portion of the TAT motif. We also did not identify 41 proteins reported in Dilks et al. (2003), which is attributed to the fact that the RRXFL[KR] motif requirement employed here is significantly more stringent than the XRRXXX used in the TATFIND program.

6.3.6 Results: Membrane-targeting transmembrane helices

Many cytoplasmic membrane proteins do not require a cleavable signal sequence in order to insert into the membrane. Instead, the presence of one or more transmembrane alpha helices and recognition by the signal recognition particle is sufficient for membrane-targeting (Bernstein, 2000; de Gier and Luirink, 2001; de Gier et al., 1998). Of the proteins without a predicted cleavable N-terminal signal peptide, 1042 (18.7%) were predicted by Phobius and TMHMM (Krogh et al., 2001) to contain at least one transmembrane helix. Some of these likely represent N-terminal uncleaved signal anchors, while in most cases it appears that internal helices may be sufficient for cytoplasmic membrane targeting.

6.4 PhoA fusion survey of the *P. aeruginosa* proteome

6.4.1 Methods

The following steps were carried out by Dr. Shawn Lewenza, a postdoctoral fellow in the laboratory of Dr. R.E.W. Hancock. *Escherichia coli* DH5α was used as the recombinant host for a *P. aeruginosa*-alkaline phosphatase (PhoA) fusion library. Genomic DNA from *P. aeruginosa* H103 was isolated, partially digested with *Sau*3AI and size fractionated on 1% agarose-Tris-acetate-EDTA (TAE) gels. Digested DNA in the size range 1 to 3 kb was excised, gel-purified and ligated into *Bam*HI digested, phosphatase-treated plasmids pJDT1, pJDT2, and pJDT3 that contain single base additions to permit the coding fragment to fuse in the correct reading frame to `*phoA* (Mdluli et al., 1995).

Ligation products were used to transform electrocompetent *E. coli* DH5α. Transformants were recovered and screened for alkaline phosphatase activity on alkaline phosphatase indicator LB agar containing 75 mM Na_2HPO_4 to repress endogenous PhoA activity, 100 ug/ml ampicillin and 90 ug/ml of the chromogenic alkaline phosphatase substrate BCIP (5-bromo-4-chloro-3-indolyl phosphate) as previously described (Bina et al., 1997). All blue colonies were picked to 96-well microtitre plates and sub-cultured in LB broth containing 50 ug/ml ampicillin.

I then carried out the remaining steps in collaboration with Dr. Lewenza. Plasmids from PhoA positive clones were purified in 96-well format (Qiagen, Mississauga, Canada), visualized for yield on 96-lane 1% agarose-TAE gels and sequenced using Big Dye Terminator chemistry (Applied Biosystems, Foster City, CA) on a Basestation 51 Fragment Analyzer (MJ Research, Waltham, MA) and a *`phoA*-specific sequencing primer directed towards the upstream cloned region as previously described (Bina et al., 1997).

The *P. aeruginosa* genes upstream of the truncated *`phoA* gene were mapped to the *P. aeruginosa* PAO1 genome sequence by BLASTN and BLASTX analyses (Altschul et al., 1997).

6.4.2 Results

In total, 1035 PhoA positive colonies were isolated. After growth in liquid media for plasmid DNA isolation, cultures were re-inoculated onto PhoA indicator agar to examine the stability of the PhoA phenotype. In contrast to colonies transferred multiple times on solid media, cultures grown in liquid media had highly unstable alkaline phosphatase phenotypes. It was reasoned that growth in liquid media strongly selects for mutations that limit the amount of PhoA fusion protein expressed, due to the toxicity of membrane-localized PhoA fusion proteins (Manoil and Traxler, 1995) or due to high expression levels. The plasmid used has a high copy number and contains the *lac* promoter upstream of the *`phoA* gene (Mdluli et al., 1995). Cloned *P. aeruginosa* fragments that produce successful PhoA fusions may also contain strong *Pseudomonas* promoters, thus it is possible that the PhoA fusions may be expressed from either the *lac* promoter, the *Pseudomonas* promoter, or off of both.

Plasmids were purified from all PhoA positive colonies regardless of the stability of the PhoA phenotype. Only plasmids with high and intermediate yields were used as templates in sequencing reactions. In some cases, extremely low plasmid yields were observed, suggesting that plasmid loss had occurred. A total of 646 plasmids were sequenced and mapped to the *P. aeruginosa* genome to identify the gene randomly cloned upstream of the `*phoA* gene. This analysis yielded a total of 474 proteins cloned in the correct orientation to produce a PhoA fusion protein while the majority of remaining sequences were of poor quality and did not produce high scoring BLAST hits to the PAO1 genome. Eliminating the redundant BLAST hits reduced the list to 310 unique *P. aeruginosa*-PhoA fusion proteins.

The ability of these proteins to direct PhoA to the cytoplasmic membrane is likely due to the presence of an export signal. Of the 310 proteins identified in the PhoA screen, 296 displayed a predicted cleavable N-terminal signal peptide or contained one or more predicted transmembrane helices, as summarized in Table 6.2.

Table 6.2: Predicted export signals in 310 PhoA fusion proteins.

Type of export signal	Number	% of export signals
Type I	169	57.1
Possible type I	1	0.3
Type II (lipoprotein)	31	10.5
Type IV(prepilin)	5	1.7
TAT	1	0.3
Transmembrane helix	89	30.1
Total with export signals	296	-
No export signal	14	-

These data indicate that our consensus computational prediction strategy displayed high recall – in other words, a low number of false negative results was encountered. However, one cannot comment on the precision, or false positive rate, of the strategy, as the PhoA fusion simply indicates export and does not provide information on the nature of the export signal itself.

Export signals were annotated as type I, probable type I, type II, type IV, TAT, or transmembrane helix based on the predictions generated in the initial computational screen. Similar proportions of type II and type IV signal peptides were observed in both the whole genome predictions and among the PhoA fusion proteins, indicating that the prediction of these two types of signals might be relatively straightforward, particularly when compared to prediction of type I signal peptides. The proportion of type I signal peptides identified in the PhoA screen was almost 20% higher than the proportion predicted genome-wide, indicating that the predictive methods may be missing some non-canonical N-terminal signal peptides or that the PhoA fusion method preferentially identifies type I signal peptides.

There were 14 proteins identified in the PhoA screen which lacked a predicted export signal. These proteins may possess non-canonical export signals not identified by the methods used, or else incorrect assignment of start sites may have caused their true export signals to be missed. Alternatively, they may represent false positives associated with the PhoA screen. To explore these possibilities, these proteins were selected for further examination.

The protein sequences and upstream regions were examined for alternative start sites by manual scanning and GeneMark 2.4 analysis (Lukashin and Borodovsky, 1998). Ten proteins displayed potential alternative translation products, which were re-analyzed by the four signal peptide prediction methods. The alternative translation products associated with three proteins (PA0667, PA0259, PA3348) were predicted by all four methods to be exported by a cleavable type I signal peptide.

Next, the remaining proteins were compared to the ePSORTdb database of proteins of experimentally verified localization (Gardy et al., 2003) using a BLASTp search and an E-value cutoff of 1e-10. None of the proteins were homologous to exported proteins but four proteins (PA2451, PA3919, PA4091, PA2744) showed similarity to cytoplasmic proteins. Lastly, the protein sequences were similarly compared to the cPSORTdb database of proteins with computationally predicted subcellular localizations. Five proteins showed similarity to predicted cytoplasmic proteins (PA1389, PA3357, PA3673, PA4124, PA4577) and none showed similarity to PSORTdb predicted exported proteins. Of the two remaining proteins, PA1531 is similar to the periplasmic protein of

ABC transporters and PA5044 (PilM) is similar to the actin-like protein MreB from *E. coli* and *Bacillus* (Mattick, 2002). PilM is involved in type IV pili biosynthesis and is necessary for twitching motility (Mattick, 2002). These two proteins have no predicted export signal but may contain a unique export signal not identified by our methods. Thus, of the 14 PhoA fusions with no apparent export signal, only five appear to be candidates for export while the remaining nine proteins are probable false positives.

Among the proteins that produced active PhoA fusions, 150 are annotated as hypothetical or conserved hypothetical proteins. This finding suggests that these proteins are likely localized to the membrane and thus provides some preliminary information regarding the function of these proteins that have not as yet been characterized.

6.5 Conclusions

In the first part of our analysis, a genome-wide computational screen for exported proteins was performed. Multiple predictive methods, including machine learning methods and manual pattern matching, were used to identify *P. aeruginosa* PAO1 proteins containing possible export signals. In large-scale genome studies, it is critical to employ a consensus approach in order to reduce the number of false positives and to increase the confidence of the prediction. Our study reports almost 100% agreement between the genome-wide predictions and the experimental PhoA fusion data.

The consensus prediction method used here indicates that 38% of the genome encodes proteins exported via five types of export signal: type I signal peptides, type II (lipoprotein) signal peptides, type IV (prepilin) signal peptides, TAT (twin arginine transporter) signal peptides, and membrane-targeting transmembrane helices. Approximately 40% of these predicted exported proteins appear to utilize cleavable type I signal peptides, according to 3 or more of the predictive methods. A small number of false positives were observed which include 4 of the 719 proteins with strongly predicted signal peptides (PA2003, PA2554, PA3883, PA5389) that show significant similarity to cytoplasmic proteins and 3 proteins with weakly predicted signal peptides (PA1949, PA1998, PA2267). Furthermore, predictions may also reflect biases in the programs' training data, such that certain non-canonical type I signal peptides might be missed. This is likely the case with many of the 57 proteins with weakly predicted type I signal peptides since many of them appear to be candidates for export based on their annotated functions, however at least 2 of the 4 methods failed to predict a signal peptide.

The methods used to identify the other classes of signal peptide are more specialized and appear to result in better predictions when compared to type I signal peptide methods. LipoP v.1.0 predicted 185 potential lipoproteins in the genome, 109 more than are presently annotated in the pseudomonas.com database. The Pseudomonas Genome Database annotations were calculated using PSORT I, and the increase in predicted lipoproteins reported here illustrates the importance of using up to date computational methods.

We predicted 23 putative prepilin peptidase substrates in the PAO1 genome, of which six represent novel candidates. These include five proteins occurring in a cluster, four of which are annotated as probable type II secretion system proteins and may represent novel type II secretion proteins, similar to the Xcp and Hxc machinery. There are likely more prepilin peptidase substrates within the *P. aeruginosa* genome, which could be identified through searching with a more degenerate motif. For example, the supposedly invariant Gly residue preceding the cleavage site appears to be replaceable by an Ala residue, as seen in the previously identified FimT protein, as well as in PA2672 and PA2674 reported here.

Fifteen putative TAT substrates were initially identified in the present analysis, which utilized stringent criteria including the presence of a predicted N-terminal signal peptide and a match of at least 5 of the 6 residues in the RRXFL[KR] motif. An expanded analysis searching for potential TAT substrates across the whole genome – not just within the subset of proteins with predicted signal peptides – identified a further 12 possible TAT substrates. This indicates that it is important not to overlook proteins without a predicted signal peptide, as they may contain functioning TAT-directing motifs. In fact, of the 18 previously identified TAT substrates reported by Ochsner et al. (2002), only 11 contain predicted type I signal peptides. Of the 57 putative substrates reported by Dilks et al. (2003), just 28 are predicted to contain a type I signal peptide. Overall, 27 potential substrates were identified, 10 of which have not been previously described. However, there are likely many more TAT substrates within the genome with more degenerate motifs, since our prediction strategy was unable to identify the known TAT substrates phospholipase PlcH and the ferripyoverdine receptor FpvA (Ochsner et al., 2002).

Our computational analysis showed that the majority of exported proteins are likely cytoplasmic membrane proteins that lack cleavable signal peptides but possess one or more transmembrane helices as an export signal. This estimate of proteins possessing transmembrane helices, 18.7% of the *P. aeruginosa* proteome, is similar to the 18.5% of proteins predicted by PSORTb to be localized to the cytoplasmic membrane. This may reflect the fact that computational prediction of transmembrane helices is generally regarded to be more accurate than the prediction of targeting signals, due to the sequence constraints associated with crossing a lipid bilayer. As signal peptide prediction methods improve, an increase in the number of predicted exported proteins is also expected.

PhoA fusion methods are a versatile genetic tool to identify proteins that are translocated across the cytoplasmic membrane. Using a plasmid-based '*phoA* screen, 310 unique *P. aeruginosa* fusion proteins were successfully identified. This approach to identify membrane-localized proteins is as efficient as reported in previous *P. aeruginosa* membrane proteomic studies (Guina et al., 2003; Nouwens et al., 2000). A significant disadvantage of this approach is the apparent toxicity of PhoA fusion proteins, which often selects for mutations that lead to a loss of PhoA activity or even the loss of plasmid. Furthermore, PhoA fusion analysis is of limited utility in the identification of proteins using certain export systems. Proteins secreted to the extracellular space by the type I or type III systems typically lack N-terminal signal peptides and, in the case of type I ABC transporter substrates instead rely on a C-terminal secretion signal (Mackman et al., 1987; Delepelaire and Wandersman, 1990; Letoffe and Wandersman, 1992). Such proteins will not be identified through PhoA fusions. Proteins using the TAT system represent a more complex case. TAT substrates are folded in the cytoplasm prior to entering the TAT transporter. PhoA, however, is folded in the periplasm, where the necessary disulfide bonds are formed. Interestingly, one protein with a predicted TAT motif did produce an active PhoA fusion. While this could represent a false positive prediction of a TAT export signal, it may also indicate that the TAT transporter is capable of translocating an unfolded substrate, or that an active PhoA fusion can be formed in the cytoplasm.

Although there is a possibility that certain *P. aeruginosa*-specific export signals may not be recognized as PhoA fusions expressed in a recombinant *E. coli* host, the strong conservation of the inner membrane targeting and translocation machinery should not affect the export of most PhoA fusion proteins. In addition, this approach has been used previously to identify secreted proteins in *Helicobacter pylori* (Bina et al., 1997).

The 310 proteins identified in this PhoA screen reflect many of the known functions associated with the cell envelope. The outer membrane proteins identified includes those that function as porins, iron uptake receptors and efflux channels, the three largest families of outer membrane proteins, and proteins involved in secretion and adhesion. Periplasmic proteins identified included the binding components of ABC transporters, cell wall biosynthesis enzymes, stress response proteases and chaperones. Inner membrane proteins included transport proteins, chemotaxis transducers, two-component sensors, efflux pumps, cell wall biosynthesis enzymes and proteins involved in secretion.

The PhoA fusion data provided confirmation that 14% of our predicted exported proteins are indeed exported, although the export signals themselves cannot be identified. The laboratory analysis also identified 14 proteins with no predicted export signal. Three of these 14 contained mispredicted start sites and the new translation products displayed type I signal peptides. Nine of the remaining proteins showed significant similarity to known and predicted cytoplasmic proteins and likely represent false positives (3%) associated with the PhoA fusion technique. A second class of false positives, not counted in the 310 successful fusions, occurred at a similar rate and included fusions to genes in the opposite orientation of the `*phoA* gene. The false positives found in the PhoA screen do not overlap with the false positives found in the computational screen and, had the computational screen not been performed, would have gone unnoticed. The remaining proteins without a predicted signal peptide exhibited significant similarity to an exported protein and a bacterial cytoskeletal protein, however their export signals remain unclear.

In summary, a combined computational and experimental approachwas employed to identify exported proteins in *P. aeruginosa*. The approach illustrates the effectiveness of using two complementary methods for genome-wide analyses. Computational techniques have the advantage of yielding a large number of predictions – ideal for genome-wide studies – and when a consensus method is employed the number of false positive results is reduced. Laboratory methods, though they generally provide fewer results, can both confirm predictions and reveal interesting cases meriting closer inspection, including erroneous annotations and potentially unusual sequence features. This combined analytical approach is readily adaptable to other bacteria – the increase of the breadth of training data available means that current export signal predictive methods can be applied to a diverse range of organisms with accurate results, and the PhoA fusion technique is commonly used to study exported proteins.

In addition to creating the *P. aeruginosa* signal peptide dataset described in this report (Appendix D), this analysis has provided laboratory-based experimental evidence to confirm the export of 14% of the predicted export candidates, as well as the existence of 150 proteins annotated as hypothetical or conserved hypothetical. The genome-wide identification of exported proteins will help define this important subset of the *P. aeruginosa* genome, and may assist in the discovery of novel drug targets.

7 ANALYSIS II: COMPARISON OF HIGH-THROUGHPUT PROTEOMIC AND COMPUTATIONAL METHODS FOR PROTEIN SUBCELLULAR LOCALIZATION IDENTIFICATION

Portions of this chapter have been previously published in the article "Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria", co-authored by J.L. Gardy, S. Rey, and F.S.L. Brinkman in BMC Genomics, Volume 6 © 2005 Rey et al; licensed by BioMed Central Ltd.

7.1 Summary

Subcellular fractionation combined with 2D gel-based proteomics permits the identification of large numbers of proteins from distinct bacterial compartments. However, the fractionation of a complex structure like the cell into individual compartments is not a trivial task. It was hypothesized that PSORTb v.2.0 could be used as a complement to fractionation-based analyses to facilitate more accurate genome-wide analysis of protein localization. Thus a comparison of computational localization prediction methods with laboratory proteomics approaches was undertaken in order to identify the most effective current approach for genome-wide localization characterization and annotation. PSORTb version 2.0 was used to computationally predict the localization of proteins reported in ten subcellular proteome analyses of bacterial compartments, and these computational predictions were then compared to the localizations determined by the proteomics study. By using a combined approach, a number of contaminants and proteins with dual localizations were identified, and we were

able to more accurately identify membrane subproteomes. Our results allowed us to estimate the precision level of laboratory subproteome studies and it was shown that, on average, recent high-precision computational methods such as PSORTb now have a lower error rate than high-throughput laboratory methods. We note that analysis of all cellular fractions collectively is required to effectively provide localization information from laboratory studies, and we propose an overall approach to genome-wide subcellular localization characterization capitalizing on the complementary nature of recent laboratory and computational methods.

7.2 Rationale

Several types of laboratory methods are frequently used in the experimental determination of a protein's localization. Techniques such as immunofluorescence and immunoelectron microscopy (Kumar et al., 2000), PhoA fusions (Manoil and Beckwith, 1985), fluorescent-protein tagging (Chalfie et al., 1994), and Western blotting with SDS-PAGE are often applied to the analysis of either single proteins or a small sets of proteins. While such methods may provide high-quality localization information, they can be costly and/or time-consuming as compared to computational methods, and the number of proteins for which a localization site can be assigned per experiment is relatively low.

Recent developments in proteomics technology now permit experimental verification of localization in a high-throughput fashion. Techniques such as two-dimensional gel electrophoresis and mass spectrometry (Dutt and Lee, 2000; Lay, 2001; Jonsson, 2001; Peng and Gygi, 2001; Govorun and Archakov, 2002) have been frequently employed in the study of a variety of bacterial genomes, including *Pseudomonas aeruginosa* (Nouwens at al., 2002) and *Bacillus sp.* (Antelmann et al., 2001). Many of these studies have focused on distinct cellular compartments through the analysis of samples obtained by subcellular fractionation (Molloy et al., 2000; Molloy et al., 2001; Bumann et al., 2002; Huang et al., 2002; Murakami et al., 2002). A major disadvantage of these subproteome analyses is that the fractionation of a complex structure like the cell into several subcellular compartments is not a trivial task. Contamination from other cellular compartments may occur and some proteins are known to span multiple localization sites (Chung et al., 1998). Despite these limitations, however, genome-scale techniques are rapid, cost-effective, and capable of returning results for hundreds or even thousands of proteins in a single analysis.

After the development of PSORTb v.2.0, we hypothesized that by combining both high-throughput laboratory methods and computational prediction, some of the errors – particularly the potential for contamination – inherent in laboratory subproteome studies could be reduced. The existence of the high-precision PSORTb tool also raised the question of how well a computational method would compare, in terms of precision, to the laboratory methods presently available. In genome-scale analyses, do laboratory and computational methods behave equally, or are particular localization sites best predicted by one or both approaches? We therefore undertook a comparison of selected bacterial subproteomic studies with PSORTb-derived computationally predicted localization sites.

7.3 Methods

7.3.1 Selection of subproteomic studies

Ten studies were selected spanning all five localization sites in the Gram-negative bacteria *Escherichia coli* (Dukan et al., 1998; Molloy et al., 2000), *Helicobacter pylori* (Bumann et al., 2002), *Klebsiella pneumoniae* (Molloy et al., 2001), *Porphyromonas gingivalis* (Murakami et al., 2002), *Pseudomonas aeruginosa* (Nouwens et al., 2002), *Salmonella typhimurium* (Molloy et al., 2001) and *Synechocystis* (Fulda et al., 20000; Huang et al., 2002). In addition, seven supplementary Gram-positive studies were evaluated to a lesser degree to ensure that the results were generally applicable to all bacteria: the cytoplasmic fractions of *Corynebacterium glutamicum* (Hermann et al., 2001; Schaffer et al., 2001) and *Mycobacterium leprae* (Marques et al., 2004), the cytoplasmic membrane fractions of *Bacillus anthracis* (Chitlaru et al., 2004), *Mycobacterium leprae* (Marques et al., 2004) and *Mycobacterium tuberculosis* (Sinha et al., 2002) and the extracellular fractions of *Bacillus sp.* (Antelmann et al., 2001) and *Staphlycoccus aureus* (Ziebandt et al., 2001).

The vast majority of the studies used fractionation followed by two-dimensional gel electrophoresis in their analysis. Proteins were then subjected to peptide mass fingerprinting (PMF) identification. One study (Murakami et al., 2002) used fractionation followed by two successive one-dimensional SDS-PAGE electrophoresis analyses, with subsequent N-terminal amino acid sequence analysis.

7.3.2 Protein selection

For each study, we examined the proteins identified and described by the authors to ensure they met two criteria. First, the protein must have been identified through direct comparison of the spot to the sequence of the bacterial genome under study and not to a related organism. For example, in the *S. typhimurium* outer membrane study of Molloy et al. only the proteins identified by a peptide mass fingerprinting search against the *S. typhimurium* genome were selected, while proteins identified by a search against other organisms were not included. Second, the protein reported in the study had to match a GenBank record in order to retrieve the correct amino acid sequence. After these two filtering steps were applied, the final dataset consisted of 405 proteins for the Gram-negative organisms and 269 Gram-positive bacterial proteins.

7.3.3 Computational analysis

Computational predictions of localization were performed using the standalone version of PSORTb v.2.0 (Gardy et al., 2005). Proteins predicted to reside at multiple localization sites were manually identified from the PSORTb score distribution; a protein was annotated with dual localizations if PSORTb returned two sites with scores between 4.50 and 7.49 or if the SCL-BLAST module returned significant similarity to a protein known to have dual localizations. Additional limited computational analyses were performed with Proteome Analyst (Lu et al., 2004).

7.4 Results

7.4.1 Comparison of computational and laboratory-based predicted localizations

A matrix comparing the PSORTb predicted localization sites to the localizations assigned by subproteome analysis is presented in Table 7.1, together with the estimated % agreement and % coverage for each study. Full results are available in Appendix E.

Table 7.1: PSORTb v.2.0 predicted localization sites for 405 proteins reported in ten subproteome studies.

		Subproteome Study: *Organism* Localization (Number of proteins)									
		***E. coli* C (23)**	***Synechocystis* CM (63)**	***Synechocystis* P (57)**	***K. pneumoniae* OM (3)**	***S. typhimurium* OM (11)**	***E. coli* OM (39)**	***P. gingivalis* OM (6)**	***P. aeruginosa* OM (33)**	***P. aeruginosa* EC (150)**	***H. pylori* EC (20)**
PSORTb Predicted Localization	**C**	19	13	2		2	3		4	33	3
	C/CM		2								
	C/P			1							
	CM		5						1	5	
	CM/P		1				1			1	
	P	1	6	8			3			33	2
	P/OM								1		
	OM		5	3	3	6	22	2	22	9	4
	OM/EC		1			1	1		2		1
	EC						1	1	1	6	1
	Unknown	3	30	43		2	8	3	2	63	9
Comparison (%)	**Agreement**	95.0	24.2	64.3	100.0	77.8	74.2	66.7	80.6	6.9	18.2
	Coverage	87.0	52.4	24.6	100.0	81.8	79.5	50.0	93.9	58.0	55.0

Agreement is defined by *a/b*, where *a* represents the number of proteins from fraction X predicted to be resident at X or X/Y localization sites and *b* represents the total number of proteins from fraction X predicted as not unknown. Coverage is defined is *b/c*, where *c* is the total number of proteins from fraction X.

Because PSORTb is designed with an emphasis on high precision, the program returns a prediction of "unknown" if not enough information is available to make a confident prediction. 163 of the 405 proteins being compared, or 40.2%, returned a result of unknown and were not considered in the downstream analyses. Of the remaining 242 proteins, the experimentally observed localization site agreed with the computationally predicted localization site in only 104 cases, for a total % agreement of 43.0%. This figure dropped to 25.7% if the unknown proteins were included in the calculation. The figures vary significantly from study to study, with % agreement ranging from a low of 6.9% (4.0% if including unknowns) in the largest study, to a high of 100% in the smallest study. However, it is clear that among the 405 proteins, there are likely a significant number of false positives and false negatives.

7.4.2 Identification of potential contaminants

Subcellular fractionation is a widely-used method for isolating the proteins resident at a specific cellular compartment (Pasquali et al., 1999). However, a significant limitation of the technique is the problem of cross-contamination, in which small amounts of proteins from neighbouring compartments contaminate the fraction of interest. This leads to the inclusion of false positives in the resulting datasets.

With the computational and subproteomic localizations differing for as many as 93.1% of the proteins for a particular analysis, we suspected that certain subproteome studies we analyzed were prone to cross-contamination. The two studies examining the extracellular fraction, in particular, displayed agreement with the computational predications of only 6.9% and 18.2%, therefore we suspected that contamination may have been a particular problem for these studies. This may be due in part to autolysis, a process common in many bacterial species known to release cellular proteins into the extracellular milieu (Morse, 1978). It may also be due to cellular lysis during the centrifugation of the cells [19]. If we exclude the study with 100% agreement, which involves only a small (n = 3) number of proteins, we observe that the study with the most agreement between the two methods involved an analysis of the *E. coli* cytoplasm. The single possible contaminant observed in this study suggests that the cytoplasm is the easiest compartment to isolate in a subfractionation analysis.

When a number of subproteome studies of Gram-positive bacteria were analyzed, we observed a similar trend. Of the seven studies we examined, the *Corynebacterium glutamicum* and *Mycobacterium leprae* cytoplasmic experiments displayed the lowest levels of observed/predicted disagreement, at 0% and 8% respectively. However, when two Gram-positive extracellular fractions were analyzed (*Staphylococcus aureus* and *Bacillus sp.*), the % disagreement was measured at 53% and 33% – figures which are significantly lower than those observed for Gram-negative bacteria.

We next proceeded to examine the 138 disagreeing cases on an individual basis to identify the source of potential false positive results. While many false positive results appeared simply to be the result of "leaky" subfractionation, we did observe a number of cases in which a protein resident in the fraction of interest was identified along with its interacting partners from neighbouring cellular compartments. For example, Molloy et al. report the presence of the acriflavine resistance protein A (AcrA) in the outer membrane fraction however, AcrA – which is predicted by PSORTb to be a cytoplasmic membrane protein – is known to be dually localized in both the cytoplasmic membrane and the periplasm (Kawabe et al., 2000; Zgurskaya and Nikaido, 2000). AcrA interacts with the outer membrane protein TolC to form an export system, thus we suspect that AcrA was found in the outer membrane fraction due to its tight association with TolC.

Another instance of "co-fractionation by association" was observed with the PilJ protein isolated from the P. aeruginosa outer membrane fraction. This protein is predicted by PSORTb to be localized to the cytoplasmic membrane and displays significant similarity to the known cytoplasmic membrane protein methyl-accepting chemotaxis protein II from Salmonella typhimurium (Milburn et al., 1991). PilJ is part of the chemosensory systems of P. aeruginosa (Darzins, 1994), and it was likely co-fractionated through its association with another component of the chemosensory system present in the outer membrane.

We also observed several conflicting cases amongst the results when closely related proteins were examined. 85 of the 405 proteins in the analysis can be grouped into 36 groups of proteins which appear multiple times in the results. These 36 groups consist of: 1) a single protein identified more than once in the studies (*e.g.* OprE, identified in both the *P. aeruginosa* outer membrane and extracellular fractions); 2) two or more paralogs (*e.g. Synechocystis* CcmK homolog 1 and CcmK homolog 2, both identified in the cytoplasmic membrane fraction); or 3) two or more orthologs (*e.g. Helicobacter pylori* carbonic anhydrase, identified in the extracellular fraction, and *Synechocystis* carbonic anhydrase, identified in the periplasmic fraction).

We would expect these groups of closely related proteins to be isolated from the same subcellular fractions, since subcellular localization is highly conserved across diverse taxonomic lineages (Nair and Rost, 2002b). However, this is only the case for 18 of the 36 groups, although 33 of the 36 are predicted by PSORTb to reside in the same localization. Fifteen groups contain related proteins isolated from two different fractions. Two groups (the ATP synthase beta chain proteins and the elongation factor family) contain proteins isolated from three fractions, and one group (the GroEL, GroEL2 and GroES chaperonin proteins) was isolated from four different subcellular fractions. These latter three groups illustrate an important trend with respect to contamination – certain abundant, predominantly cytoplasmic, proteins are repeatedly found in the list of potential contaminants, either due to the subfractionation process or their association (even if temporary) with proteins of another localization (for example, the protein folding chaperones). In the majority of these studies, however, they are not noted as potential contaminants/co-purifying proteins.

Our analysis of false positives reveals that the potential for contamination appears to be lowest when the cytoplasm is the subfraction of interest, and highest when the extracellular fraction is analyzed. The data highlight the fact that employing a computational contaminant screening procedure is a valuable addition to a subproteome analysis. It is especially critical for extracellular analyses, as both autolysis and mechanical lysis of cells during subfractionation can release the contents of other cellular compartments into this fraction of interest. The ubiquitous cytoplasmic proteins ATP synthase beta, elongation factors, and the GroEL/ES chaperonins are frequently observed contaminants; however, many of the studies in which these proteins were identified do not address this fact. While these proteins might immediately raise a flag to most proteomics researchers, they are not commonly noted and so may not be appreciated by genomics researchers using localization data for genome annotation or cell surface drug target identification. Failure to note these proteins as potential contaminants/co-purifying proteins may also have significant consequences for bioinformatics software development. For example, inaccurate subcellular localization assignments could be propagated if the data were used as training data for a machine learning method by researchers unfamiliar with the field.

7.4.3 An estimation of the precision of subproteome 2D gel analyses

An interesting figure results from the analysis of the 44 proteins that were both isolated in a subproteome study and are present in the ePSORTdb database of proteins of known subcellular localization. In 12 of these 44 cases, the fraction from which these proteins were isolated in the subproteomic studies did not match the previously reported experimentally verified localization. If we view these 44 proteins found in ePSORTdb as "100% precise predictions", we arrive at a "true" potential contamination rate of 27.3%. Nine of these conflicting results were found in the extracellular fraction in the subproteomic experiments and may represent by-products of cellular lysis. The remaining three proteins were isolated from the *E. coli* outer membrane fraction, though they were previously shown to be periplasmic proteins. The authors of this subproteome study propose that these proteins were extracted through their association with outer membrane components, rather than improper fractionation technique.

We then carried out a more liberal analysis by investigating the 138 cases where the PSORTb and subproteomic localizations differed. For each of the 138 proteins, we attempted to determine the most probable actual localization site. Localizations for twelve proteins, mentioned above, were found in ePSORTdb. We next looked for a published report of localization in the literature for the remaining 126 proteins. If no published information was available, we then looked for significant (E > 1e-10) similarity to a protein of known localization.

In this fashion, we were able to confirm that the localization predicted by PSORTb was correct in 87 of the 138 proteins. For the remaining 51 proteins, neither published localization information nor similarity to a protein of known localization was observed, and we were unable to determine whether the PSORTb or subproteomic localization site was correct. The results of this analysis are presented in Table 7.2.

Table 7.2: Estimation of subproteome study error rate.

	Subproteome Study: *Organism* Localization										
	E. coli C	*Synechocystis* CM	*Synechocystis* P	*K. pneumoniae* OM	*S. typhimurium* OM	*E. coli* OM	*P. gingivalis* OM	*P. aeruginosa* OM	*P. aeruginosa* EC	*H. pylori* EC	Total
Proteins identified	23	63	57	3	11	39	6	33	150	20	405
Disagreements	1	25	5	0	2	8	1	6	81	9	138
PSORTb errors	0	0	0	0	0	0	0	0	2	1	3
Laboratory errors	0	4	1	0	2	6	1	3	36	5	58
Error %	0.0	6.3	1.8	0.0	18.2	15.4	16.7	9.1	24.0	25.0	14.3

Disgreement is defined as the number of proteins from fraction X predicted NOT to be resident at X or X/Y localization sites. PSORTb errors are defined as the number of disagreeing cases for which the literature confirmed PSORTb's prediction was incorrect. . Laboratory errors are defined as the number of disagreeing cases for which the literature confirmed the subproteome study's assignment of localization was incorrect. Error % is defined as the number of laboratory errors divided by the total number of proteins identified.

Using this more liberal analysis, we estimated the average error rate of laboratory subproteome experiments to be 14.3%. Estimated error rate values varied considerably between studies, from a low of 0% (in the *K. pneumoniae* outer membrane analysis, in which only 3 proteins were investigated) to a high of 25.0% (in the *H. pylori* study of the extracellular fraction). Again, we observed that extracellular studies appeared to have the highest error rates due to the strong potential for contamination discussed earlier. On average, though, the subproteomic analysis error rate for all localizations was significantly higher than the error rate of 4% previously reported for PSORTb (Gardy et al., 2005).

7.4.4 Reducing information loss: proteins with dual localization sites

A second disadvantage of subcellular fractionation is the associated information loss. Certain proteins have domains in two or more neighbouring cellular compartments, some may cleave into two products, each residing at a different site (Henderson et al., 2000), and others may be found at different localizations over time, or during different environmental conditions (Hefty et al., 2002). Because subproteome studies typically address a single cellular compartment, it is quite difficult to identify multiply-localized proteins from the results.

Computational methods can help to reduce the information loss associated with subproteome studies. When a disagreement is observed in cases where the computational and subproteomic localization sites are neighbours, it may indicate a dually localized protein. An example found in the present analysis is the ATP synthase AtpG (beta prime subunit). This protein was extracted from the *Synechocystis* cytoplasmic membrane fraction but was predicted as a cytoplasmic protein by PSORTb. Inspection of the literature reveals that AtpG contains domains located in both the cytoplasm and cytoplasmic membrane (Takeyasu et al., 1996; Dunn et al., 2000; Dunn et al., 2001).

PSORTb also flags proteins predicted to reside in two compartments. Thirteen of the 405 proteins are predicted to reside at dual localization sites, with the bulk of these predicted as outer membrane/extracellular. This particular combination of localization sites suggests an autotransporter – a protein with a beta-barrel transporter domain and extracellular globular domain that is cleaved and released after translocating through the pore formed by the transporter domain. Indeed, many of the 13 proteins flagged by PSORTb are known autotransporters, including esterase and the *H. pylori* vacuolating cytotoxin.

Although PSORTb can assist in the identification of dually-localized proteins, false negatives are still possible. If the observed site and the single predicted sites are identical, a protein's secondary localization will still go undetected. Though it may not always be feasible, a potential solution to this problem would be to perform 2D gel analyses of all five compartments in one experiment. Not only would this aid in the identification of proteins with multiple localization sites, a comparison of the amounts of protein present in each fraction could be of use when screening for potential contamination.

7.4.5 Comparison of PSORTb with previously reported contaminant screening procedures

Our results illustrate that it is important to screen the results of a subproteome study for potential errors. However, many groups do not perform such a screen, or employ approaches which are limited in their utility.

The authors of two of the subproteomic studies analyzed here performed basic contaminant screens. In the *Synechocystis* cytoplasmic membrane study, the 63 proteins identified were submitted to TMHMM (Krogh et al., 2001). Seventeen of these proteins were classified as integral membrane proteins based on the presence of one or more helices. The remaining 46 were annotated as peripherally-associated membrane proteins and were then analyzed by

SignalP (Bendtsen et al., 2004). Proteins with predicted signal peptides were classified as associated to the periplasmic face of the membrane, while those without predicted signal peptides were classified as peripherally associated to the cytoplasmic face.

Using only a single localization predictive method such as TMHMM to identify a feature often results in false positives, particularly in alpha helix detection, where signal peptides are often mistaken for helices. Furthermore, by describing the proteins with no detected helices as peripherally membrane-associated, there is a failure to recognize the fact that these proteins may represent potential contaminants from other fractions. Had PSORTb been used as a screening tool, the authors would have been able to identify 22 potential errors amongst their results with a relatively high degree of confidence.

The authors of the *E. coli* outer membrane study compared the SWISS-PROT localization site for the proteins they identified to the amounts of those proteins detected on the 2D-gel. They reported that, with the exception of the flagellin protein, only proteins annotated as integral outer membrane proteins were detected in significant levels. They posit that the remaining proteins, detected at lower levels, may exhibit a functional association with proteins in the outer membrane. However, this explanation does not account for several potential cytoplasmic or cytoplasmic membrane contaminants, such as the dihydrolipoamide succinyltransferase SucB (Knapp et al., 1998, 2000), which were also isolated. A screen such as this also has the potential to produce a high number of false negatives – outer membrane proteins present in low quantities which are mistaken for potential contaminants.

While the authors of the two studies mentioned above do not claim that their approaches identify all contaminants, we found that a robust and comprehensive method such as PSORTb outperforms single methods designed to analyze specific features, such as signal peptides or transmembrane helices. This is not surprising, as it has long been recognized that multi-component approaches to prediction achieve the best performance. Though dually localized proteins likely represent only a small fraction of proteins in the cell, they often represent interesting biological cases, including proteins that play pivotal roles in antimicrobial resistance (i.e. efflux proteins (Poole et al., 1993), and virulence (i.e. BrkA (Fernandez and Weiss, 1994)) and thus should not be overlooked.

7.4.6 Optimal identification of cytoplasmic membrane proteins requires a combined computational and laboratory approach

Examining the detailed PSORTb results for the proteins reviewed in the present analysis, we observed an interesting trend in the output of the HMMTOP module, which predicts the number of transmembrane alpha-helices in a query protein. Of the 405 proteins analyzed by HMMTOP, only six proteins contained three or more predicted helices. Even more surprising was that only three of these six were identified in the *Synechocystis* cytoplasmic membrane study. When three cytoplasmic membrane subproteome studies in Gram-positive bacteria were analyzed, the same trend was observed, with only six out of 269, or 2.2%, of proteins predicted to contain three or more transmembrane helices

(TMHs). We then analyzed the complete *Synechocystis* proteome with PSORTb, predicting a total of 540 cytoplasmic membrane proteins, of which 461 contain three or more transmembrane helices.

Our results indicate that 2D gel electrophoresis of the cytoplasmic membrane fraction is only capable of identifying a small proportion of the multi-pass membrane proteins in a given proteome, likely due to the low pI and poor solubility of these proteins (Santoni et al., 2000). While other techniques can be used to identify these proteins in the laboratory – for example, liquid chromatography coupled with tandem mass spectrometry and affinity labelling (Goshe et al., 2003; Blonder et al., 2004) – PSORTb is a cheaper and faster solution which is capable of identifying these proteins with a high degree of precision.

While PSORTb appears to outperform laboratory subproteomic methods for the identification of proteins with three or more transmembrane helices, the opposite is true for membrane-associated proteins with one or two helices. In their analysis of the *Synechocystis* cytoplasmic membrane fraction, the authors of the study report 40 membrane-associated proteins. PSORTb, on the other hand, only confidently identifies three such dually localized proteins – two with cytoplasmic domains, and one with a periplasmic domain. In order to maintain a high level of precision, PSORTb requires that one of the following criteria be met to identify a cytoplasmic membrane protein: three or more predicted TMHs, similarity to a known membrane protein, or a positive result from the cytoplasmic membrane SVM module. As a result of these stringent criteria, a large number of cytoplasmic membrane-associated proteins with one or two helices are not identified by PSORTb.

Our observations indicate that the cytoplasmic membrane presents a special case for both laboratory and computational analysis. If a true picture of the membrane proteome is desired, it is necessary to use a combined approach, in which a computational method is used to identify integral cytoplasmic membrane proteins, while a laboratory method is used to identify cytoplasmic membrane-associated proteins.

7.5 Conclusions

7.5.1 Comparing the precision of laboratory and computational methods

In the present analysis, we compared the localizations predicted by the computational method PSORTb to the localizations of 405 proteins reported in ten subproteome 2D gel electrophoresis studies. The data generated in our analysis indicates that subproteome studies vary greatly in terms of their precision. Certain small studies of particular fractions, such as the analysis of three *K. pneumoniae* outer membrane proteins or 23 *E. coli* cytoplasmic proteins, display low or non-existent apparent error rates. Larger studies and those focusing on particular localizations – including the extracellular milieu – can contain significant levels of false positive, or contaminant proteins.

We attempted to estimate the precision associated with subproteome studies using two approaches. In the first, more stringent approach, a comparison of 44 proteins against the ePSORTdb database of proteins of experimentally verified localization yielded a rough estimate of false positives of 27.3%. A second approach, in which we attempted to determine the true localization of 138 proteins using literature and homology-based approaches, yielded an estimate of 14.3%.

While our approximate error rate is by no means a definitive estimate and was not calculated using large samples, it does illustrate the importance of evaluating the results of a subproteome study with a critical eye. While errors associated with each study do vary, on average as many as 1 out of every 4–7 results could be erroneous.

Even more notable is the observation that while our estimated precision of subproteome analysis exceeds that of early predictive tools such as PSORT I (with a reported precision of 59.6%), current high-precision computational methods such as PSORTb (with 96% precision) appear to outperform laboratory subproteome studies, generating fewer false positive results. While it is true that measured precision values calculated from cross-validation studies of test datasets represent a slight overestimation of precision, even a more conservative estimate of 90% precision still exceeds the levels attained by most high-throughput laboratory methods. In other words, PSORTb, first released in 2003, appears to be the first computational method developed that outperforms high-throughput laboratory studies for localization prediction. Other computational methods have since been developed that also have high accuracy and slightly higher recall, such as Proteome Analyst. However, no method has yet been developed that is as precise as PSORTb.

7.5.2 Limitations of computational methods

While our comparison of the precision achieved by computational and laboratory subproteome analyses indicates that certain predictive tools have surpassed wet-bench methods for localization identification, there are a number of caveats associated with the use of computational tools.

Of the 405 proteins submitted to PSORTb, only 59.8% returned a predicted localization site and in only 43% of these cases did the predicted site match the observed site. The 40.2% "unknown" rate we observed is well below the recall of 82% reported in the paper describing PSORTb. Such a discrepancy between "practical" values and "theoretical" values is frequently observed with machine learning methods, due to the fact that the data used to train and test the method is generally quite well-annotated while "real world" data, on the other hand, contains large numbers of hypothetical proteins.

Unfortunately, until machine-learning methods – including PSORTb – are trained on much larger datasets, the gap between recall values is not likely to improve significantly. In the interim, we recommend that users employ additional predictive strategies with higher recall values. Proteome Analyst uses a different approach to PSORTb in generating its predictions – keywords are extracted from SWISS-PROT annotations of proteins homologous to a given query; these keywords are then passed to a machine learning classifier. Proteome Analyst displays excellent precision – the authors report an overall precision of 95.9% for Gram-negative bacteria – and although its coverage when applied to whole genomes is generally comparable to PSORTb, it did provide a much larger number of predictions for the dataset analyzed here – of the 405 proteins submitted, Proteome Analyst returned a predicted localization site or sites for 398.

The performance of a given method can also vary significantly depending on the organism being analyzed. For example, PSORTb was able to generate predictions for only 25% of the proteins identified in the *Synechocystis* periplasmic fraction. Several factors may explain this low rate of coverage, including particularities of the morphology of *Synechocystis* sp., the low number of *Synechocystis* proteins included in PSORTb's training dataset, and the fact that three-quarters of the proteins found in the periplasmic fraction are annotated as hypothetical proteins. This is in contrast to the excellent coverage achieved by PSORTb in the analysis of the *E. coli* cytoplasmic fraction, which reflects the fact that as a model organism, *E. coli* proteins occur frequently in PSORTb's training data.

A method's performance also varies between localization sites and, in general, correlates with the amount of training data available for a given localization. PSORTb performs very well when identifying both cytoplasmic and outer membrane proteins, but is not able to make as many predictions for periplasmic and extracellular proteins. Proteins resident at specific localization sites – for example, the periplasm and the extracellular space – can be similar to the point that differentiating the two based on sequence alone can be difficult.

It is also important to note that every predictive method will generate a certain number of false positive results, and that it is critical to keep the measured precision of a given method in mind when carrying out a computational analysis. For example, some computational methods, such as CELLO, have a measured precision of only 71.5%.

7.5.3 Limitations of laboratory methods

Laboratory analyses also carry with them a number of caveats. We have already shown that one of the major disadvantages of subproteomic studies is the potential for contamination via leaky fractionation or lysis. Growth conditions can also affect the results of a subproteome study. Different growth conditions can alter the expression of a particular protein, thus while a subproteome study can provide valuable data about expression under a given condition, they may not yield a global picture of the proteins expressed by a bacterium. The parameters of the experiment can also play a key role in determining which proteins are identified from a gel.

It is critical to choose an appropriate pH gradient for maximum resolution of total proteins, and even then standard methods may not detect or separate low abundance or hydrophobic proteins. Protein complexes can also be problematic if their subunits are difficult to disassociate (Santoni et al., 2000; Cordwell et al., 2001).

7.5.4 Proposed method for the optimal characterization of cellular compartments

We have shown that computational and laboratory-based analyses of specific cellular compartments complement each other, with each method contributing to improve the accuracy of the other. Although both methods do display certain limitations, each offers a number of significant advantages, which we have summarized in Table 7.3. In order to capitalize on these advantages, we propose that genome-scale studies aimed at cataloguing the proteins of a particular cellular compartment adopt a complementary approach in which both methods are used.

Table 7.3: Advantages and disadvantages of computational and subproteomic approaches to localization analysis.

Computational methods	**Subproteomics analysis**
Advantages	
Rapid predictions	Provides condition-specific information
Detailed information about sequence features	Confirms expression of hypothetical ORFs
Can identify potential contaminants	Large-scale source of localization data for proteins lacking homologs
Can identify hydrophobic integral membrane proteins	
Disadvantages	
Lower performance on "non-model" organisms	Time/resource consumption
Lower performance for localizations with small amount of training data	Low abundance and hydrophobic proteins often missed
Not condition-specific	Contamination
	Difficult to identify multiply localized proteins

With respect to the subproteomic aspect of such a study, we suggest that rather than analyze a single cellular compartment, a study ought to analyze all available compartments. By determining the relative abundance of a protein in each compartment, a researcher will able to quickly flag potential contaminants and identify proteins with complex localization profiles – dual localizations or localization that varies temporally.

After retrieving the set of protein sequences corresponding to the spots on a 2D gel, the proteins should be submitted to a high-precision localization prediction method for analysis. PSORTb is the most precise localization prediction tool available, and its consensus approach allows the user to acquire detailed information about protein features, such as homology to a protein of known localization, or the presence of a signal peptide, transmembrane helices, or specific sequence motifs and patterns. Proteome Analyst is a second high-precision method which complements PSORTb well through the use of an annotation-based approach.

The computationally predicted and experimentally observed localization sites should then be compared. In cases where the computational and laboratory methods disagree, detailed analysis of the individual protein should be carried out. Through examination of the literature and further computational analysis, very often a confident call regarding the protein's true localization can be made. An excellent model is provided by Elias et al. (2005), who employ a multi-faceted approach – including PSORT I, PSORTb, and in-depth examination of individual proteins – to the analysis of their results from a study of *Shewanella oneidensis* hypothetical proteins.

The combination of 2D gel analysis and PSORTb prediction can provide a remarkably clear and genome-scale picture of protein localization in a given bacterium. Of course, these methods are no replacement for the hypothesisdriven detailed investigation of individual proteins. Instead, they provide an accurate jumping-off point for the in-depth analysis of specific proteins using additional techniques. As both computational and laboratory high-throughput approaches improve in terms of both precision and recall, however, we see an increasingly important role for these methods in the fields of molecular biology and genomics.

8 ANALYSIS III: COMPARATIVE GENOMICS ANALYSIS OF BACTERIAL PROTEIN LOCALIZATION: IMPLICATIONS FOR NETWORK EVOLUTION

8.1 Summary

The development of the high precision localization prediction method PSORTb permitted a global analysis of protein localization across multiple bacterial genomes. We examined the percentage of proteins predicted to be resident at each localization site for 236 sequenced bacterial genomes and observed several notable trends. These include: an increase in cytoplasmic proteins in thermophilic and hyperthermophilic bacteria, an increased proportion of cell surface proteins in pathogenic bacteria, and a general trend in which the proportion of a proteome at each localization site is generally well-conserved across species, regardless of genome size. This latter observation may reflect a method of adaptive evolution in which new functions are gained through the acquisition of "peripheral subnetworks" – small subnetworks of genes whose products are functionally related and span multiple cellular compartments.

8.2 Rationale

With the development of PSORTb v.2.0 and our study indicating that its precision has surpassed that of high-throughput laboratory methods, the comparative analysis of localization suddenly became feasible.

While researchers have previously examined certain localization sites – cytoplasmic proteins and exported proteins, for example – across a range of genomes, no one has yet examined localization in a global context – looking at all localization sites at one time. With over 200 completely sequenced bacterial genomes available, this represents a rich source of data yet to be mined.

We set out to examine protein localization in sequenced bacterial genomes using PSORTb, hypothesizing that such a global analysis would not only provide information about differences in localization in certain bacteria, but would also yield answers regarding the evolutionary history of protein localization.

8.3 Methods

We analyzed the October 3, 2005 release of cPSORTdb, comprising 689,275 proteins from the genomes of 162 Gram-negative and 74 Gram-positive organisms. For each of the organisms, the percentage of proteins resident at each cellular compartment was calculated by dividing the number of proteins predicted at each localization site by the total number of proteins encoded in the genome. Complete data is provided in Appendix F.

8.4 Results

8.4.1 An estimate of the percentage of proteins at each bacterial localization site

The arithmetic mean, minimum and maximum percentage of proteins at each localization site are summarized in Table 8.1.

Table 8.1: A summary of the percentage of the proteome resident at each cellular compartment for 162 Gram-negative and 74 Gram-positive bacteria.

		Percentage		
Gram stain	Localization site	Arithmetic Mean	Minimum	Maximum
Negative	Cytoplasm	33.2	18.8	56.0
	Cytoplasmic membrane	16.9	11.2	24.2
	Periplasm	1. 6	0	3.8
	Outer membrane	2.5	0.6	11.0
	Extracellular	0.4	0.1	1.1
	Multiple	2.0	0.8	6.1
	Unknown	43.4	21.2	58.6
Positive	Cytoplasm	50.7	41.7	59.9
	Cytoplasmic membrane	19.9	14.0	24.5
	Cell wall	0.9	0	2.2
	Extracellular	3.0	1.0	7.3
	Multiple	0.8	0.2	2.7
	Unknown	24.8	16.8	33.9

The majority of proteins encoded by a genome are predicted to be cytoplasmic. In Gram-positive bacteria, we estimate the mean proportion of cytoplasmic proteins to be 50.7%. Our calculation of the mean for Gram-negative bacteria, however, is significantly lower, at 33.2%. This reflects the fact that the Gram-negative version of PSORTb – specifically the SVM-based module for the identification of cytoplasmic proteins – exhibits comparatively low recall when applied to this class. We propose that the actual average proportion of cytoplasmic proteins in Gram-negative bacteria is closer to the 50% observed in Gram-positive organisms. Our data support the earlier conclusions of Schatz and Dobberstein (1996), who estimate the cytoplasmic fraction of an average cell to be approximately 50%.

Although the calculated mean proportions of predicted cytoplasmic proteins are quite different between Gram-negative and Gram-positive bacteria, the maximum observed proportions are similar, at 56% and 59.9%, respectively. This indicates an approximate upper bound for the proportion of predicted cytoplasmic proteins at 60%. Recognizing that PSORTb and other predictive algorithms do not display perfect recall and thus underestimate true biological proportions, we propose that the upper bound for cytoplasmic proteins *in vivo* is likely higher.

Cytoplasmic membrane proteins are the second most prevalent in the bacterial cell. Because methods for the *in silico* identification of this class of protein exhibit both high precision and high recall, we observe much less discrepancy in the estimated mean proportions between Gram-negative and Gram-positive organisms; our results indicate a mean proportion of 16.9% in the former and 19.9% in the latter. Again, a consistent upper bound to proportions is observed, this time at approximately 24%.

Our data is in agreement with that of Bendtsen et al. (2005), who report cytoplasmic membrane proportions of 15-20% across 217 bacterial genomes. Granseth et al. (2005), however, report higher average proportions of 22-26% across 204 bacterial genomes using an improved method taking into account laboratory-derived topology information.

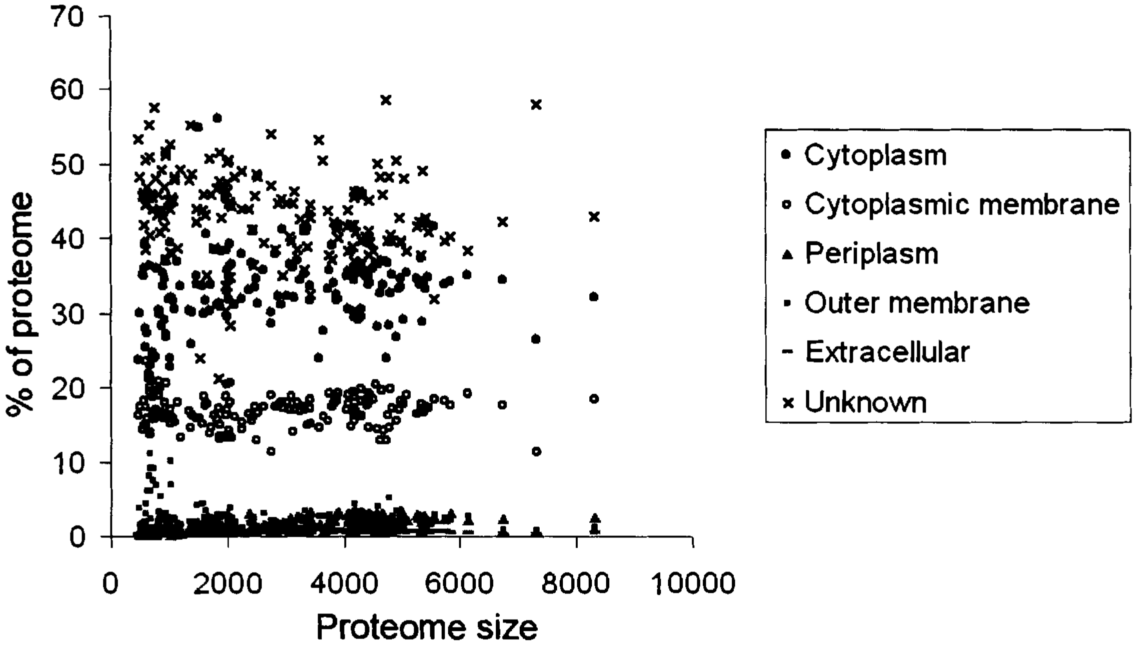
Taken together, our results and those of earlier groups indicate that roughly 50-70% of a bacterial genome is devoted to cytoplasmic and cytoplasmic membrane proteins, though this proportion may be as high as 80% in certain organisms, such as *Thermoanaerobacter tengcongensis*. The remaining proportion of proteins require some sort of export signal – signal peptide or otherwise – to be secreted beyond the cytoplasmic membrane to their final localization site. Predicted proportions of exported proteins, including cell wall, periplasmic, outer membrane and extracellular proteins, are highest in the Mycoplasmae, consistent with the earlier predictions of Saleh et al. (2001) and Schneider (1999) and reflective of these organisms' atypical cell envelopes, which lack a peptidoglycan cell wall and do not contain a periplasm. Certain Bacilli, including *Staphyloccocus* sp. and *Bacillus cereus*, also appear to be prolific exporters. It is important to note, however, that the reported recall of PSORTb and other predictive methods is lower for exported proteins – particularly periplasmic and extracellular proteins – than for cytoplasmic and cytoplasmic membrane proteins. Thus the mean proportions for these fractions as described above is likely an underestimation.

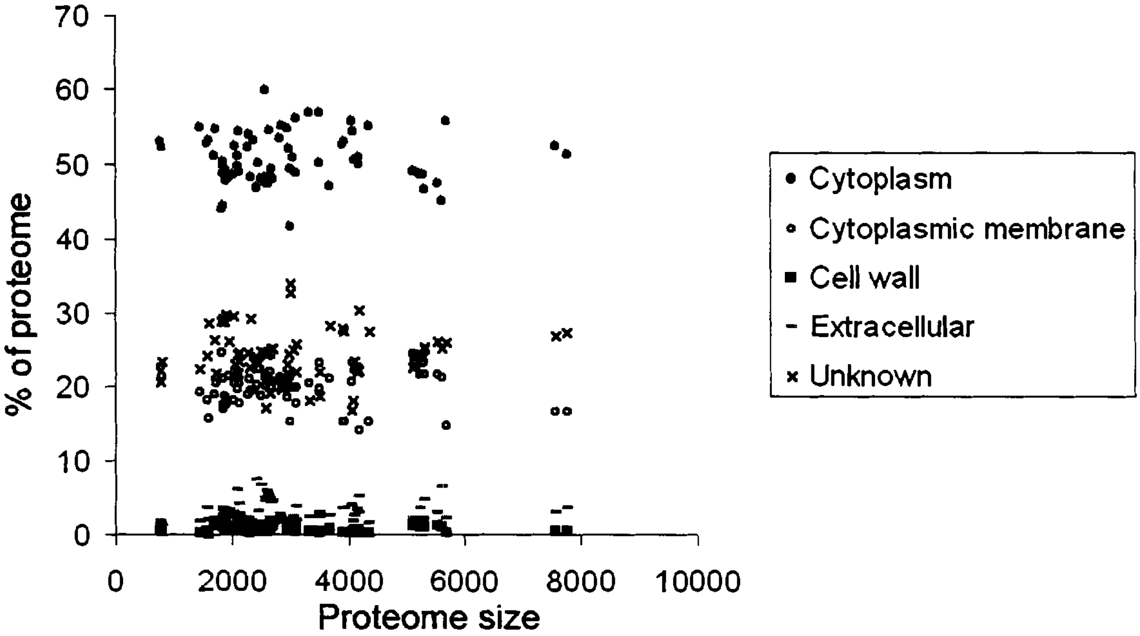
8.4.2 Proportions of proteins at each localization site are consistent across species regardless of genome size or lifestyle

A scatter plot of the proportions of proteins at each localization site illustrates that, with the exception of the Mycoplasmae, values are generally consistent across organisms (Figure 8.1) and that neither lifestyle nor genome size has an appreciable impact on the proportions at any one site. Indeed, 71% of the data points fall within one standard deviation of the mean, while 96% fall within two standard deviations. While Bendtsen et al. (2005) and Granseth et al. (2005) have previously described such consistency in the proportions of cytoplasmic membrane proteins, ours are the first reports illustrating that this trend extends to all localization sites.

Figure 8.1: **Percentages of proteins at each localization site are generally well-conserved across bacterial species.**

Top: 162 Gram-negative bacteria. Bottom: 74 Gram-positive bacteria. Organisms are arranged along the x-axis by proteome size (# of proteins).



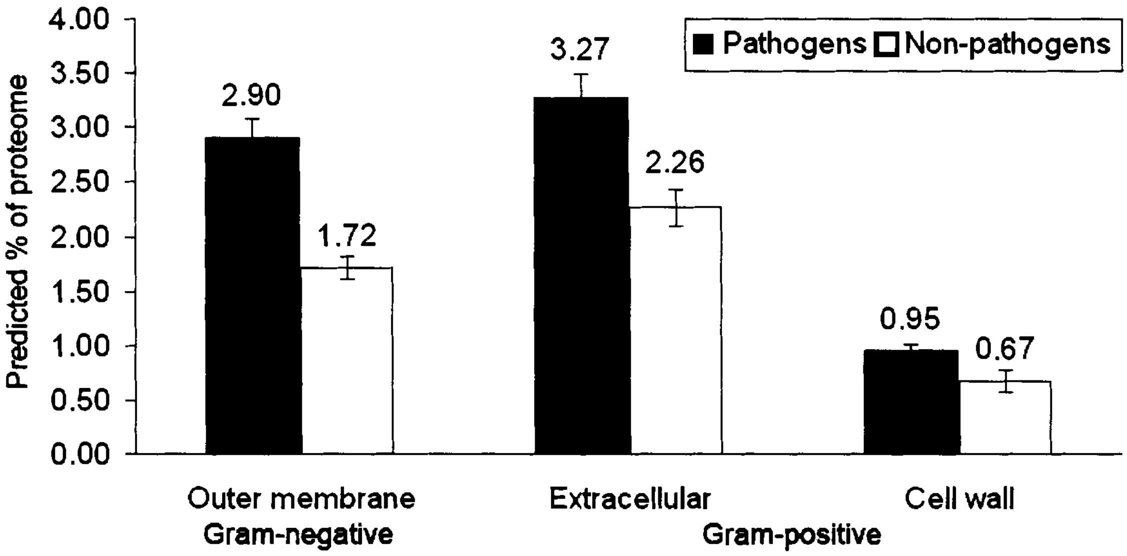


8.4.3 Conservation of proportions reflects peripheral subnetwork-based adaptive evolution

The consistency inherent in the proportions of proteins at each localization site has implications for network-based evolution. Our observations suggest that the number of proteins in each cellular compartment changes in concert rather than independently – an increase or decrease in the number of proteins at one site must be accompanied by a similar change in the number of proteins at other localization sites. This in turn may imply that bacterial evolution is the result of the simultaneous acquisition or loss of multiple genes rather than singletons (Lawrence and Roth, 1996, Boucher et al., 2003). Recent observations extend this hypothesis, noting that groups of genes aquired by a bacterium display physiologically coupled functions – typically related to the early stages of a particular metabolic pathway – and are likely acquired as "subnetworks" which attach to existing networks within a cell (Pal et al., 2005a). We propose, therefore, that adaptive evolution in bacteria requires the acquisition of a subnetwork of genes whose products are not only functionally related – but also span multiple subcellular localization sites.

In an analysis of *E. coli* metabolic networks, Pal et al. (2005a) also propose that protein components of subnetworks mediating adaptive evolution exhibit "peripheral" functions – they are active in the early stages of metabolic pathways and are engaged in specialized processes including nutrient uptake and early reaction steps. They then attach to existing cellular networks at a more "central" point in the pathway – for example, a protein involved in later stages of the pathway exhibiting a more generalized function. Our data indicates that the protein subcellular localizations associated with these subnetworks are also peripheral. We have also noted that the mean proportions of peripherally localized proteins differ between pathogenic and non-pathogenic bacteria (Figure 8.2). The increases in peripherally localized proteins in pathogens not only provides support for the peripheral subnetworks hypothesis but also may reflect the pathogens' need for antigenic diversity in their complement of surface-exposed proteins.

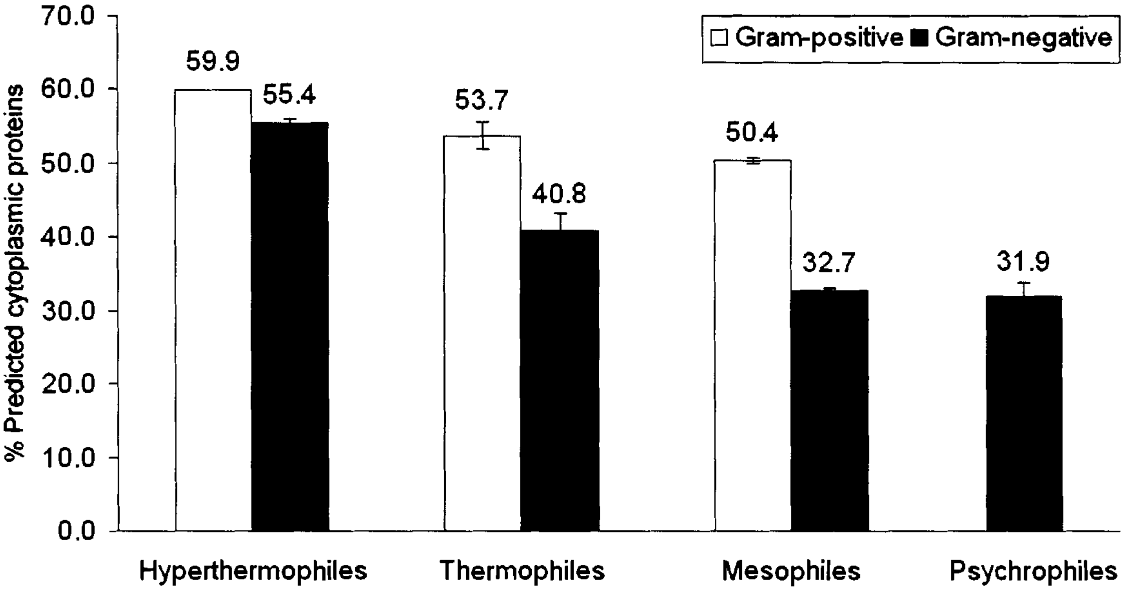
Figure 8.2: Mean proportions of peripherally localized proteins differ between pathogenic and non-pathogenic bacteria.



8.4.4 Thermophilic and hyperthermophilic bacteria display elevated proportions of cytoplasmic proteins

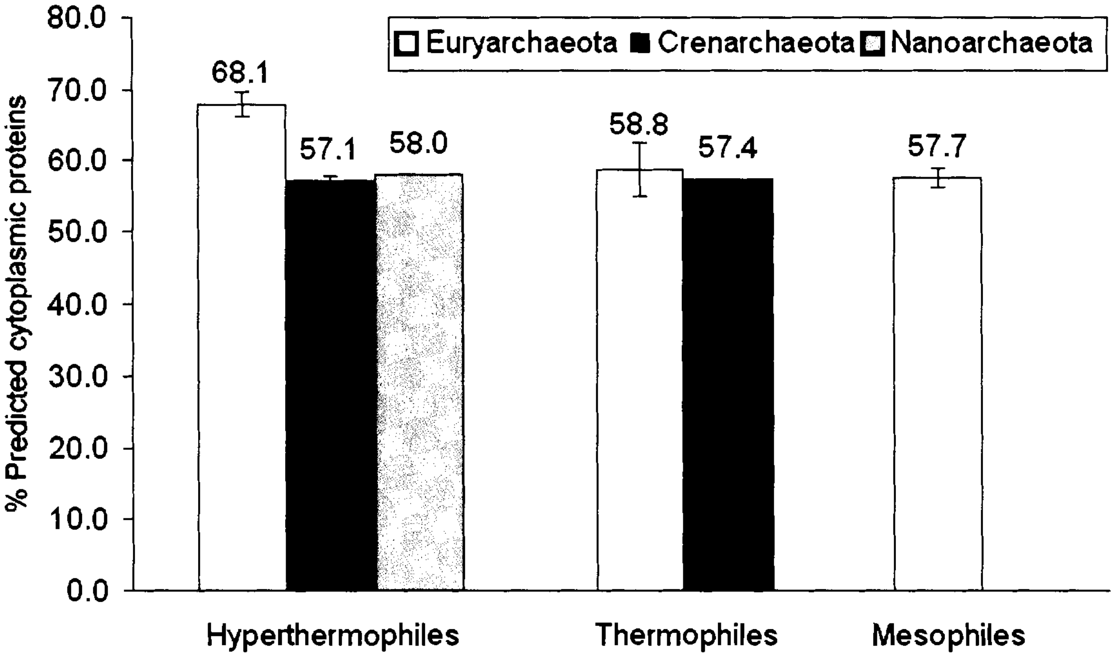
Although proportions are generally well-conserved across species, certain small variations, such as the increase in cell surface proteins in pathogens, are present. One of the most interesting variations was observed when bacteria were grouped according to optimal temperature range. A comparison of the mean proportions across groups revealed that thermophilic and hyperthermophilic bacteria displayed elevated proportions of cytoplasmic proteins relative to mesophiles and psychrophiles (Figure 8.3).

Figure 8.3: Mean proportions of cytoplasmic proteins are elevated in thermophilic and hyperthermophilic bacteria.



Noting that the hyperthermophiles in particular showed an increase in cytoplasmic proteins and that these organisms tended to occur at the base of the tree of life, we extended the analysis to include 24 archaeal genomes in order to examine the trend from an evolutionary perspective. To date, a high-precision localization prediction method for archaea has yet to be developed, although an archaeal-specific release of PSORTb is planned. Instead, we analyzed the archaeal genomes using the Gram-positive version of PSORTb. In this analysis a similar, yet phylogenetically restricted, trend was observed (Figure 8.4).

Figure 8.4: Mean proportions of cytoplasmic proteins are elevated in hyperthermophilic Euryarchaeaota.



Hyperthermophilic Euryarchaeota were predicted to contain over 10% more cytoplasmic proteins than other archaea, including mesophiles, thermophiles and hyperthermophilic Crenarchaeota. While ours are the first analyses to illustrate the increase in cytoplasmic proteins in hyperthermophilic bacteria and archaea, other groups have previously noted low proportions of exported proteins in archaea (Schneider, 1999; Saleh et al., 2001).

Several explanations for this possibly significant increase in cytoplasmic proteins in the hyperthermophiles are possible. Hyperthermophilic organisms may be biased towards low proportions of exported or surface-exposed proteins, which would be prone to denaturation in the extreme environments they favour. An increased proportion of cytoplasmic proteins may have been characteristic of the ancestral state of the first primitive cell since the hyperthermophiles examined are also disproportionately basal-branching organisms (for example, *Aquifex aeolicus* and *Thermotoga maritima*; Klenk et al, 2004). Sequencing of representative genomes of the early-diverging Korarchaeota (Barns et al., 1996) and subsequent analysis of their cytoplasmic proteins would further our understanding of the ancestral basis of this trend.

8.5 Conclusions

By applying the high-precision localization method PSORTb to the genomes of 236 bacteria, we have estimated the proportions of proteins resident at each cellular compartment in bacteria. The majority of a genome is predicted to encode cytoplasmic and cytoplasmic membrane proteins. Methods for the computational identification of these classes of proteins are relatively well-developed thanks to large amounts of training data in the case of the former and extensive research into the physiological character of the latter. However, the remainder of the genome encodes exported proteins, which currently present a challenge to predictive methods. Recall for these proteins is low, resulting in a likely underestimation of the exported fraction of a cell. To develop a clearer picture of the true biological distribution of proteins throughout the cell, both an increase in the amount of training data for underrepresented sites such as the periplasm as well as improvements in the predictive methods themselves will be required.

When the proportions of proteins in each cellular compartment were visualized in a graph (Figure 8.1), it became clear that the values are relatively consistent, regardless of genome size or lifestyle. This suggests to us that in order to maintain this balance, the acquisition of proteins resident at one compartment is coupled to similar changes in the protein complement of neighbouring compartments. We propose that these changes occur through the acquisition of "peripheral subnetworks" – groups of proteins spanning multiple localizations that attach to existing pathways within the recipient cell.

Recent work on adaptive evolution in bacteria has provided insights into how this process works (Pal et al, 2005a, Pal et al., 2005b, Light et al, 2005). Our data further extend the hypothesis, and allow us to propose a model for the subnetwork-driven evolutionary process.

In our model, peripheral subnetworks represent a primary unit of adaptive evolution. Upon acquisition, either horizontally or through gene duplication, a subnetwork will integrate with an existing pathway at a highly connected, centrally localized point. Acquisition of the subnetwork typically confers a new function upon the recipient organism, enabling rapid adaptation to a new environment or stimulus (for example, the ability to uptake and metabolize a different carbon compound). Pathways whose topology has been preserved throughout evolution and whose functional and structural interactions are uniquely co-evolved (Shi et al., 2005) are not disrupted through the assimilation of peripheral nodes, and the packaging of a novel gene with its functional partners increases the chances that a newly acquired function will successfully establish itself and become vertically inherited. Indeed, Pal et al. confirm that integration of a newly acquired network is six times more likely to be successful if a physiologically coupled downstream component is already present in the recipient genome (2005b).

APPENDICES

Appendix A

PROSITE motifs used in the PSORTb motifs modules:

PS00538 Cytoplasmic Membrane
RTE[EQ]Q.{2}[SA][LIVM].[EQ]TAASMEQLTATV

PS00755 Cytoplasmic Membrane
[GST][LIVMF][LIVMFCA].[LIVMF][GSA][LIVM].P[LIVMFY]{2}.[AS][GSTQ][LIVMFAT]{3}[EQ][LIVM
FA]{2}

PS00756 Cytoplasmic Membrane
[LIVMFYW]{2}.[DE].[LIVM][STDNQ].{2,3}[GK][LIVMF][GST][NST]G.[GST][LIV][LIVFP]

PS00192 Cytoplasmic Membrane
[DENQ]...G[FYWMQ].[LIVMF]R..H

PS00193 Cytoplasmic Membrane
P[DE]W[FY][LFY]{2}

PS01303 Cytoplasmic Membrane
[GSDN]WT[LIVM].[FY]W.WW

PS00449 Cytoplasmic Membrane
[STAGN].[STAG][LIVMF]RL.[SAGV]N[LIVMT]

PS00713 Cytoplasmic Membrane
P.{0,1}G[DE].[LIVMF]{2}.[LIVM]{2}[KREQ][LIVM]{3}.P

PS00714 Cytoplasmic Membrane
P.G.[STA].[NT][LIVMC]DG[STAN].[LIVM][FY].{2}[LIVM].{2}[LIVM][FY][LI][SA]Q

PS00217 Cytoplasmic Membrane
[LIVMF].G[LIVMFA]..G.{8}[LIFY]..[EQ].{6}[RK]

PS00218 Cytoplasmic Membrane
[STAGC]G[PAG].{2,3}[LIVMFYWA]{2}.[LIVMFYW].[LIVMFWSTAGC]{2}[STAGC]...[LIVMFYWT].[
LIVMST]...[LIVMCTA][GA]E.{5}[PSAL]

PS00943 Cytoplasmic Membrane
N...[DEH]..[LIMF]D..[VM].R[ST]..R.{4}G

PS00077 Cytoplasmic Membrane
[YWG][LIVFYWTA]{2}[VGS]H[LNP].V.{44,47}HH

PS00221 Cytoplasmic Membrane
[HNQA].NP[STA][LIVMF][ST][LIVMF][GSTAFY]

PS00428 Cytoplasmic Membrane
[NV].{5}[GTR][LIVMA].P[PTLIVM].G[LIVM]...[LIVMFW][LIVMFW]S[YSA]GG[STN][SA]

PS00994 Cytoplasmic Membrane
R[LIVM][GSA]EV[GSA]ARF[STAIV]LD[GSA][LM]PGKQM[GSA]ID[GSA][DA]

PS0r0896 Cytoplasmic Membrane
G[LIVM]{2}.D[RK]LGL[RK]{2}.[LIVM]{2}W

PS00897 Cytoplasmic Membrane
P.[LIVMF]{2}NR[LIVM]G.KN[STA][LIVM]{3}

PS00942 Cytoplasmic Membrane
[QEK][RF]G.{3}[GSA][LIVF][WL][NS].[SA][HM]N[LIV][GA]G

PS01307 Cytoplasmic Membrane
A[LMF].[GAT]T[LIVMF].G.[LIVMF].{7}P

PS00594 Cytoplasmic Membrane
IG[GA]GM[LF][SA].P.{3}[SA]G.{2}F

PS01039 Periplasmic
G[FYIL][DE][LIVMT][DE][LIVMF]...[LIVMA][VAGC]..[LIVMAGN]

PS01037 Periplasmic
[GAP][LIVMFA][STAVDN]....[GSAV][LIVMFY]{2}Y[ND]...[LIVMF].[KNDE]

PS01040 Periplasmic
[AG].{6,7}[DNEG]..[STAIVE][LIVMFYWA].[LIVMFY].[LIVM][KR][KRHDE][GDN][LIVMA][KNGSP][F
W]

PS01157 Periplasmic
GSYPSGHT

PS00635 Periplasmic
[LIVMFY][APN].[DNS][KREQ]E[STR][LIVMAR].[FYWT].[NC][LIVM]..[LIVM]P[PAS]

PS00087 Periplasmic
[GA][IMFAT]H[LIVF]H.{2}[GP][SDG].[STAGDE]

PS00123 Periplasmic
[IV].DS[GAS][GASC][GAST][GA]T

PS00332 Periplasmic
G[GN][SGA]G.R.[SGA]C.{2}[IV]

PS00401 Periplasmic
K.[NQEK][GT]G[DQ].[LIVM].{3}QS

PS00556 Periplasmic
[LIVMA]{4}C[LIVMFA]T[LIVMA]{2}.{4}[LIVM].[RG].{2}L[CY]

PS00757 Periplasmic
NPK[ST]SG.AR

PS00968 Periplasmic
[LIVFAG].[GASV][LIVFA].[IV]H.{3}[LIVM][GSTAE][STANH].{1,3}[STN]W[LIVMFYW]

PS00969 Periplasmic
[EQ].{4}H.{5}[GSTA].{3}[FY].{3}[AG].{2}[AV]H.{7}P

PS00576 Outer Membrane
[LIVMFY]..G..Y.F.K..[SN][STAV][LIVMFYW]V

PS00694 Outer Membrane
(G[LIVMFY]N[LIVM]KYRYE)

PS00695 Outer Membrane
([FYW]..G.GY[KR]F)

PS01151 Outer Membrane
[VL][PASQ][PAS]G[PAD][FY].[LI][DNQSTAP][DNH][LIVMFY]

PS00875 Outer Membrane
[GR][DEQKG][STVM][LIVMA]{3}[GA]G[LIVMFY].{11}[LIVM]P[LIVMFYWGS][LIVMF][GSAE].[LIVM
]P[LIVMFYW]{2}..[LV]F

PS00834, PS00835 Outer Membrane
((WTD.S.HP.T).*(AGYQE[ST]R[FYW]S[FYW][TN]A.GG[ST]Y))|((AGYQE[ST]R[FYW]S[FYW][TN]
A.GG[ST]Y).*(WTD.S.HP.T))

PS01068 Outer Membrane
[LIVMA].[GT].[TA][DA]..[DG][GSTP]..[LFYDE][NQS]..[LI][SG][QE][KRQE]RA..[LV]...[LIVMF].{4,5}[
LIVM]....[LIVM]...[SG].G

PS00274 Extracellular
[KT]..NW..T[DN]T

PS00330 Extracellular D.[LI].{4}G.D.[LI].GG.{3}D

GGXGXD Extracellular
(GG.G.D.*){4}

PS00369 Cytoplasmic
G[LIVM]H[STAV]R[PAS][GSTA][STAMVN]

PS00589 Cytoplasmic
[GSTADE][KREQSTIV].{4}[KRDN]S[LIVMF]{2}.[LIVM]..[LIVM][GADE]

PS00077 Cytoplasmic Membrane
[YWG][LIVFYWTA]{2}[VGS]H[LNP].V.{44,47}

PS00192 Cytoplasmic Membrane
[DENQ]...G[FYWMQ].[LIVMF]R..

PS00193 Cytoplasmic Membrane
P[DE]W[FY][LFY]{2}

PS00216 Cytoplasmic Membrane
[LIVMSTAG][LIVMFSAG]..[LIVMSA][DE].[LIVMFYWA]GR[RK].{4,6}[GSTA]

PS00217 Cytoplasmic Membrane
[LIVMF].G[LIVMFA]..G.{8}[LIFY]..[EQ].{6}[RK]

PS00218 Cytoplasmic Membrane
[STAGC]G[PAG].{2,3}[LIVMFYWA]{2}.[LIVMFYW].[LIVMFWSTAGC]{2}[STAGC]...[LIVMFYWT].[
LIVMST]...[LIVMCTA][GA]E.{5}[PSAL]

PS00449 Cytoplasmic Membrane
[STAGN].[STAG][LIVMF]RL.[SAGV]N[LIVMT]

PS00713 Cytoplasmic Membrane
P.{0,1}G[DE].[LIVMF]{2}.[LIVM]{2}[KREQ][LIVM]{3}.P

PS00714 Cytoplasmic Membrane
P.G.[STA].[NT][LIVMC]DG[STAN].[LIVM][FY]..[LIVM]..[LIVM][FY][LI][SA]Q

PS00872 Cytoplasmic Membrane
[DG]...G...[DN].{6,8}[GA][KRHQ][FSA][KR][PT][FYW][LIVMWQ][LIV].[GAFV][GSTA]

PS00943 Cytoplasmic Membrane
N...[DEH]..[LIMF]D..[VM].R[ST]..R.{4}G

PS01022 Cytoplasmic Membrane
[GA][GAS][LIVMFYWA][LIVM][GAS]D.[LIVMFYWT][LIVMFYW]G...[TAV][IV]...[GSTAV].[LIVMF]...[
GA]

PS01023 Cytoplasmic Membrane
[FYT]..[LMFY][FYV][LIVMFYWA].[IVG]N[LIVMAG]G[GSA][LIMF]

PS01219 Cytoplasmic Membrane
D[FYWS]AG[GSC].{2}[IV].{3}[SAG]{2}.{2}[SAG][LIVMF].{3}[LIVMFYWA]{2}.[GK].R

PS01303 Cytoplasmic Membrane
[GSDN]WT[LIVM].[FY]W.WW

PS01327 Cytoplasmic Membrane
[KR]GN[LIV]{2}D[LIVM]A[LIVM][GA][LIVM]{3}G

PS00755 Cytoplasmic Membrane
[GST][LIVMF][LIVMFCA].[LIVMF][GSA][LIVM].P[LIVMFY]{2}.[AS][GSTQ][LIVMFAT]{3}[EQ][LIVM
FA]{2}

PS00756 Cytoplasmic Membrane
[LIVMFYW]{2}.[DE].[LIVM][STDNQ].{2,3}[GK][LIVMF][GST][NST]G.[GST][LIV][LIVFP]

PS01072 Cellwall
[LVFYT].[DA].{2,5}[DNGSATPHY][FYWPDA].{4}[LIV]..[GTALV].{4,6}[LIVFYC]..G.[PGSTA].{2,3}[M
FYA].[PGAV].{3,10}[LIVMA][STKR][RY].[EQ].[STALIVM]

PS00274 Extracellular
[KT]..NW..T[DN]T

PS00277 Extracellular
YGG[LIV]T.{4}N

PS00278 Extracellular
K..[LIVF].{4}[LIVF]D...R..L.{5}[LIV]Y

PS00429 Extracellular
ARP...K.S.TNAYNVTT..[DN]G...YG

Appendix B

Outer membrane motifs:

AAGAAG
AAGKIS
AALAAN
AANANI
AASAVE
AASTTA
AAYRYS
ADAADR
ADLFPR
AEIREK
AELEQQ
AETLAE
AGAGAE
AGARYI
AGGAIF
AGLAAL
AGLGAA
AGQASA
AGSGQV
AGTVTT
AKVTIT
ALAAPL
ALAAVL
ALALLA
ALAQQA
ALASQA
ALAVTT
ALGALG
ALGGGW
ALKVKR
ALLPSA
ALLVAG
ALQEFG
ANAAEI
APAQAE
AQAAVE
AQTLEQ
ARIEVG
ASAREG
ASNGLR.*LGRLGL
ATGAAV
ATLGLV
ATLTLT
AVAVAL
AVDFHG
AVDVAR
AVIAEV
CFCLPL
DGQDGD
DGTLNL
DIQEFI
DIRVDG
DNSKTD
DPRVKG
DRWQST
DSVPLL
DTLVVT
DYGSLS
EAYLAL
EELGDL
EFLDRL
EGINKV
ELAQAN
ELDLFG
ELGGKR
ELSLWI
EQGLEN
ESLGLR
ESRRAL
FGDSLS
FGRSKD
FKLNYA
FMGWMW
FRDFAE
FSLKNS
FTGKGY
FVSLNA
GASAGV
GASSGY
GDGGAI
GDSLSD
GELSLS
GFIEDS
GFNLNY
GFSSRD
GGAISS
GGAIYA
GGANAA
GGGAIY
GGKGGA
GGKRGA
GGRLRA
GGVVNI
GGVWGR
GKGGAI
GLGSAA
GPFVIN
GQTVVI
GSFDYG
GSGALG
GSGGSL
GSGQLS
GTILFS
GTLSGK
GTLSSA
GTLTVS
GTVSGL
GVGINL
GVKTDL
GVLKTD
GYFDFR
HRIATL
IDNTST
IEARIV
IEQGTV
IGAARA
IGRAGL
IGVLTD
ISLTAN
ISSPRL
IYRNSP
KEVLRD
KGGAIY
KINEGP
KITINN
KLSADE
KTLFTK
KVPFLG
LAAAVA
LAEPNL
LAFAGL
LALGGL
LALSIS
LAPAQA
LATASL
LAVAVA
LDKQFF
LDLELS
LDVLDA
LFSLLE
LGAATA
LGALFR
LGDIPV
LGGDGI
LGNLFK
LGRLGL
LGTYLT
LIACLS
LIDGKP
LLAATP
LLDAQR
LLDVLD
LNLSIP
LPIFTA
LRPGMT
LSAGVS
LSERRA
LSISGN
LSLLPL
LTLDPD
LTQPLF
LTVTDT
LVAKAD
LVDGVR
LVVDLS
MKKLLP
MKKTLL
NAAFSN
NALSKR
NAQLSL
NATLNG
NEVTGL
NGTVNI
NISRNF
NNGAIL
NNGTLI
NNNINA
NQLSVS
NRSTLS
NSIYID
NTKTSS
NTTINS
NVTLQG
NYAAGG
PGVSVG
PLGLSD
PLLGDI
PTLDLT
PVLAAD
PVQVLA
QANAAT
QASWLA
QFYLGA
QGTVTL
QLGGDI
QPLFDY
QSSSAA
QTDDET
RAALLP
RADLFP
RALALA
RDFAEN
RGPEGR
RLNALE
RLNQLS
RLVSLV
RVEILR
SAGSLA
SALALA
SASRTV
SFLPSV
SGLGRA
SGQTYN
SGSFNF
SGSSSS
SLAGTV
SLIALA
SLLAGS
SLLALS
SLLDVL
SLLIGG
SLQQPL
SLSLPP
SNITGG
SQLDWK
SRFSTS
SRLTLG
SRPVAD
SSSSSSSS
STVVEL
SVNIRG
SVNVVG
TADGQL
TAPVFA
TASLLA
TATDLG
TDTPAV
TFYTKL
TGAGTL
TGDIGN
TGTLNI
TITGNK
TLGDGY
TLSGKT
TLSSAG
TMVVTA
TTLSAG
TVSLSG
TVVSAP
VAAALV
VAQRTA
VASELA
VATYRN
VDGSLS
VDGVLK
VGDSSK
VGVAFG
VGVTAK
VIQNSG
VLDAQR
VNNLFD
VPAGPF
VPGLTF
VPLLGD
VPPGPF
VPVAQV
VPWDQA
VRLDGG
VRLVVA
VRYDEA
VSGPPR
VSGRFD
VSPSSE
VSSGGT
VSVVTS
YAERGL
YQDGSA
YTVLDQ
YTVRGF

Appendix C

PROSITE profiles used in the PSORTb profile modules:

PS50862 Cytoplasmic

PS50253 Cytoplasmic Membrane

PS50283 Cytoplasmic Membrane

PS50850 Cytoplasmic Membrane

LPXTG Cell Wall

PS50830 Extracellular

Appendix D

Results of the *P. aeruginosa* genome-wide exported protein predictions and subsequent PhoA fusion screen are available online. The following files are provided:

http://www.genome.org/cgi/data/15/2/321/DC1/1
Signal peptide, cleavage site and transmembrane helix predictions for the complete *P. aeruginosa* proteome.

http://www.genome.org/cgi/data/15/2/321/DC1/6
Successful PhoA fusions.

Appendix E

PSORTb predicted localization sites for the 405 Gram-negative subproteome studies described in section 7, organized by NCBI GI number.

			Localization	
Reference	NCBI GI	Organism	Laboratory	PSORTb
Molloy et al., 2000	113930	*E. coli*	OM	Unknown
Molloy et al., 2000	114546	*E. coli*	OM	C
Dukan et al., 1998	114580	*E. coli*	C	C
Huang et al., 2002	114698	*Synechocystis sp.*	CM	C
Molloy et al., 2000	118906	*E. coli*	OM	P
Dukan et al., 1998	119191	*E. coli*	C	C
Dukan et al., 1998	119201	*E. coli*	C	C
Dukan et al., 1998	119391	*E. coli*	C	Unknown
Molloy et al., 2000	120312	*E. coli*	OM	EC
Dukan et al., 1998	123441	*E. coli*	C	C
Dukan et al., 1998	125161	*E. coli*	C	C
Molloy et al., 2000	125964	*E. coli*	OM	OM
Molloy et al., 2000	128373	*E. coli*	OM	Unknown
Molloy et al., 2000	129043	*E. coli*	OM	C
Molloy et al., 2000	129135	*E. coli*	OM	OM
Molloy et al., 2000	129146	*E. coli*	OM	OM
Molloy et al., 2000	129151	*E. coli*	OM	OM
Molloy et al., 2000	129161	*E. coli*	OM	OM
Molloy et al., 2000	129595	*E. coli*	OM	OM
Huang et al., 2002	131173	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	131190	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	131387	*Synechocystis sp.*	CM	Unknown
Dukan et al., 1998	131650	*E. coli*	C	C
Molloy et al., 2000	132480	*E. coli*	OM	Unknown
Dukan et al., 1998	133030	*E. coli*	C	C
Dukan et al., 1998	133976	*E. coli*	C	C
Dukan et al., 1998	134659	*E. coli*	C	Unknown
Dukan et al., 1998	135173	*E. coli*	C	C
Molloy et al., 2000	135980	*E. coli*	OM	OM
Molloy et al., 2000	136459	*E. coli*	OM	OM
Bumann et al., 2002	137076	*H. pylori*	EC	Unknown
Molloy et al., 2000	140430	*E. coli*	OM	OM
Dukan et al., 1998	140742	*E. coli*	C	C
Molloy et al., 2000	232021	*E. coli*	OM	Unknown
Molloy et al., 2000	399000	*E. coli*	OM	CM/P
Molloy et al., 2001	400158	*K. pneumoniae*	OM	OM
Molloy et al., 2000	416728	*E. coli*	OM	OM
Huang et al., 2002	417545	*Synechocystis sp.*	CM	Unknown
Dukan et al., 1998	543784	*E. coli*	C	C
Molloy et al., 2000	585619	*E. coli*	OM	P
Molloy et al., 2000	585997	*E. coli*	OM	Unknown
Dukan et al., 1998	586733	*E. coli*	C	C
Molloy et al., 2000	730225	*E. coli*	OM	OM
Molloy et al., 2000	730963	*E. coli*	OM	P
Molloy et al., 2001	731022	*S. typhimurium*	OM	OM
Dukan et al., 1998	1169970	*E. coli*	C	C
Molloy et al., 2000	1171903	*E. coli*	OM	OM
Nouwens et al., 2002	1172509	*P. aeruginosa*	OM	CM
Dukan et al., 1998	1176189	*E. coli*	C	Unknown
Molloy et al., 2001	1279830	*K. pneumoniae*	OM	OM
Molloy et al., 2001	1345766	*S. typhimurium*	OM	C
Murakami et al., 2002	1536824	*P. gingivalis*	OM	EC
Huang et al., 2002	1705799	*Synechocystis sp.*	CM	C

			Localization	
Reference	NCBI GI	Organism	Laboratory	PSORTb
Dukan et al., 1998	1730032	*E. coli*	C	C
Nouwens et al., 2002	2276417	*P. aeruginosa*	EC	Unknown
Huang et al., 2002	2493271	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	2493297	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	2493552	*Synechocystis sp.*	CM	C
Huang et al., 2002	2493553	*Synechocystis sp.*	CM	C
Huang et al., 2002	2494260	*Synechocystis sp.*	CM	C
Huang et al., 2002	2494772	*Synechocystis sp.*	CM	Unknown
Molloy et al., 2001	2495350	*S. typhimurium*	OM	Unknown
Molloy et al., 2000	2497721	*E. coli*	OM	Unknown
Bumann et al., 2002	2499106	*H. pylori*	EC	OM/EC
Huang et al., 2002	2500503	*Synechocystis sp.*	CM	C
Huang et al., 2002	2506210	*Synechocystis sp.*	CM	C
Bumann et al., 2002	2506418	*H. pylori*	EC	OM
Molloy et al., 2000	2506737	*E. coli*	OM	OM
Molloy et al., 2000	2506898	*E. coli*	OM	OM/EC
Huang et al., 2002	2506910	*Synechocystis sp.*	CM	Unknown
Dukan et al., 1998	2506993	*E. coli*	C	C
Dukan et al., 1998	2507064	*E. coli*	C	C
Molloy et al., 2000	2507089	*E. coli*	OM	OM
Molloy et al., 2000	2507166	*E. coli*	OM	Unknown
Molloy et al., 2000	2507462	*E. coli*	OM	OM
Molloy et al., 2000	2507463	*E. coli*	OM	OM
Molloy et al., 2000	2507464	*E. coli*	OM	OM
Molloy et al., 2000	2507465	*E. coli*	OM	OM
Murakami et al., 2002	2541865	*P. gingivalis*	OM	Unknown
Murakami et al., 2002	2827775	*P. gingivalis*	OM	Unknown
Molloy et al., 2000	2851539	*E. coli*	OM	OM
Molloy et al., 2001	2896133	*S. typhimurium*	OM	OM/EC
Molloy et al., 2000	3025033	*E. coli*	OM	Unknown
Fulda et al., 2000	3025122	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	3025123	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	3025125	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	3025187	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	3121786	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	3123076	*Synechocystis sp.*	CM	C
Dukan et al., 1998	3220012	*E. coli*	C	C
Nouwens et al., 2002	3386644	*P. aeruginosa*	EC	EC
Molloy et al., 2001	3445382	*S. typhimurium*	OM	OM
Murakami et al., 2002	3901098	*P. gingivalis*	OM	OM
Huang et al., 2002	3913624	*Synechocystis sp.*	CM	CM
Molloy et al., 2001	3914223	*K. pneumoniae*	OM	OM
Huang et al., 2002	3915410	*Synechocystis sp.*	CM	OM/EC
Molloy et al., 2000	3915413	*E. coli*	OM	OM
Dukan et al., 1998	3916024	*E. coli*	C	P
Murakami et al., 2002	5759277	*P. gingivalis*	OM	OM
Murakami et al., 2002	5759279	*P. gingivalis*	OM	Unknown
Bumann et al., 2002	6015162	*H. pylori*	EC	Unknown
Huang et al., 2002	6016607	*Synechocystis sp.*	CM	CM
Huang et al., 2002	6225604	*Synechocystis sp.*	CM	CM
Molloy et al., 2001	6625701	*S. typhimurium*	OM	Unknown
Molloy et al., 2001	6625703	*S. typhimurium*	OM	C
Fulda et al., 2000	6919991	*Synechocystis sp.*	P	Unknown
Nouwens et al., 2002	7081488	*P. aeruginosa*	EC	EC
Dukan et al., 1998	9911121	*E. coli*	C	C
Huang et al., 2002	13634042	*Synechocystis sp.*	CM	C
Nouwens et al., 2002	15595218	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595224	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595238	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15595243	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595283	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595337	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595360	*P. aeruginosa*	OM	OM

			Localization	
Reference	NCBI GI	Organism	Laboratory	PSORTb
Nouwens et al., 2002	15595480	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15595488	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15595488	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15595489	*P. aeruginosa*	OM	C
Nouwens et al., 2002	15595497	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15595498	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15595506	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595511	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15595544	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15595618	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595620	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595624	*P. aeruginosa*	OM	P/OM
Nouwens et al., 2002	15595743	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15595753	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595769	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15595799	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15595806	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15595814	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595819	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595820	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15595963	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596004	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15596009	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15596049	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15596053	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15596085	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596092	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15596140	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15596155	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15596155	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15596159	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15596169	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596170	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15596271	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596289	*P. aeruginosa*	EC	EC
Nouwens et al., 2002	15596289	*P. aeruginosa*	OM	EC
Nouwens et al., 2002	15596347	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596356	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15596363	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15596368	*P. aeruginosa*	EC	CM/P
Nouwens et al., 2002	15596375	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15596384	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15596446	*P. aeruginosa*	EC	EC
Nouwens et al., 2002	15596468	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15596484	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596485	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15596534	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596539	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596690	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15596736	*P. aeruginosa*	OM	C
Nouwens et al., 2002	15596776	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15596784	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15596801	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15596945	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15596965	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15596974	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15596974	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15596997	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15597001	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15597068	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15597142	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15597400	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15597446	*P. aeruginosa*	EC	C

			Localization	
Reference	NCBI GI	Organism	Laboratory	PSORTb
Nouwens et al., 2002	15597487	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15597790	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15597829	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15597863	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15597956	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15597956	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15598135	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598234	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15598277	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598382	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15598386	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15598423	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15598432	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598434	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598454	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15598509	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15598641	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598662	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598723	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15598725	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598731	*P. aeruginosa*	OM	OM/EC
Nouwens et al., 2002	15598807	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598823	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15598844	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15598851	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15598871	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598888	*P. aeruginosa*	EC	OM
Nouwens et al., 2002	15598919	*P. aeruginosa*	EC	EC
Nouwens et al., 2002	15598929	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598932	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15598965	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598980	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598982	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15598985	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15598995	*P. aeruginosa*	OM	Unknown
Nouwens et al., 2002	15599002	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15599008	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599014	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599031	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599053	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15599060	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15599061	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599117	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599118	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15599126	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599135	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15599248	*P. aeruginosa*	EC	CM
Nouwens et al., 2002	15599262	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15599305	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15599370	*P. aeruginosa*	EC	EC
Nouwens et al., 2002	15599399	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599403	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15599429	*P. aeruginosa*	EC	CM
Nouwens et al., 2002	15599444	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599445	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15599469	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599482	*P. aeruginosa*	OM	C
Nouwens et al., 2002	15599521	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15599566	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15599581	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15599582	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15599613	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15599629	*P. aeruginosa*	EC	Unknown

			Localization	
Reference	NCBI GI	Organism	Laboratory	PSORTb
Nouwens et al., 2002	15599656	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599691	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599692	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15599697	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15599760	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599809	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15599820	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15599856	*P. aeruginosa*	OM	Unknown
Nouwens et al., 2002	15599882	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15599930	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599937	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15599955	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15599959	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15599986	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600005	*P. aeruginosa*	EC	CM
Nouwens et al., 2002	15600115	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15600137	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600167	*P. aeruginosa*	OM	OM
Nouwens et al., 2002	15600209	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15600229	*P. aeruginosa*	EC	CM
Nouwens et al., 2002	15600230	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600239	*P. aeruginosa*	EC	CM
Nouwens et al., 2002	15600273	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15600275	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15600289	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600305	*P. aeruginosa*	OM	OM/EC
Nouwens et al., 2002	15600346	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15600360	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15600377	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600440	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15600463	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600510	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15600516	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15600523	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600562	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600571	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600608	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15600615	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15600682	*P. aeruginosa*	EC	P
Nouwens et al., 2002	15600698	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600707	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600734	*P. aeruginosa*	EC	C
Nouwens et al., 2002	15600738	*P. aeruginosa*	EC	Unknown
Nouwens et al., 2002	15600747	*P. aeruginosa*	EC	C
Bumann et al., 2002	15644804	*H. pylori*	EC	OM
Bumann et al., 2002	15644859	*H. pylori*	EC	Unknown
Bumann et al., 2002	15644995	*H. pylori*	EC	Unknown
Bumann et al., 2002	15645005	*H. pylori*	EC	OM
Bumann et al., 2002	15645443	*H. pylori*	EC	C
Bumann et al., 2002	15645522	*H. pylori*	EC	Unknown
Bumann et al., 2002	15645523	*H. pylori*	EC	Unknown
Bumann et al., 2002	15645633	*H. pylori*	EC	P
Bumann et al., 2002	15645712	*H. pylori*	EC	EC
Bumann et al., 2002	15645732	*H. pylori*	EC	P
Bumann et al., 2002	15645787	*H. pylori*	EC	C
Bumann et al., 2002	15645800	*H. pylori*	EC	Unknown
Bumann et al., 2002	15645899	*H. pylori*	EC	Unknown
Bumann et al., 2002	15646063	*H. pylori*	EC	OM
Bumann et al., 2002	15646067	*H. pylori*	EC	C
Bumann et al., 2002	15646164	*H. pylori*	EC	Unknown
Molloy et al., 2000	15832424	*E. coli*	OM	C
Molloy et al., 2000	16129736	*E. coli*	OM	OM
Huang et al., 2002	16329185	*Synechocystis sp.*	CM	P

			Localization	
Reference	NCBI GI	Organism	Laboratory	PSORTb
Huang et al., 2002	16329195	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16329198	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16329241	*Synechocystis sp.*	CM	OM
Fulda et al., 2000	16329323	*Synechocystis sp.*	P	OM
Huang et al., 2002	16329327	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16329341	*Synechocystis sp.*	CM	C/CM
Huang et al., 2002	16329361	*Synechocystis sp.*	CM	C
Fulda et al., 2000	16329372	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16329387	*Synechocystis sp.*	P	P
Fulda et al., 2000	16329409	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16329434	*Synechocystis sp.*	CM	P
Fulda et al., 2000	16329573	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16329577	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16329600	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16329650	*Synechocystis sp.*	P	P
Huang et al., 2002	16329661	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16329662	*Synechocystis sp.*	CM	CM
Fulda et al., 2000	16329708	*Synechocystis sp.*	P	OM
Huang et al., 2002	16329721	*Synechocystis sp.*	CM	OM
Fulda et al., 2000	16329725	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16329729	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16329729	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16329841	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16329947	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16329967	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16330041	*Synechocystis sp.*	CM	OM
Huang et al., 2002	16330090	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16330090	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16330095	*Synechocystis sp.*	CM	C
Fulda et al., 2000	16330142	*Synechocystis sp.*	P	C
Fulda et al., 2000	16330225	*Synechocystis sp.*	P	P
Fulda et al., 2000	16330228	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330232	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330233	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330236	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330237	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330280	*Synechocystis sp.*	P	C/P
Huang et al., 2002	16330287	*Synechocystis sp.*	CM	C/CM
Fulda et al., 2000	16330319	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16330332	*Synechocystis sp.*	CM	P
Fulda et al., 2000	16330376	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330406	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16330412	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16330444	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16330486	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16330605	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16330613	*Synechocystis sp.*	CM	C
Fulda et al., 2000	16330681	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16330711	*Synechocystis sp.*	CM	CM
Fulda et al., 2000	16330843	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330845	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16330863	*Synechocystis sp.*	CM	OM
Fulda et al., 2000	16330863	*Synechocystis sp.*	P	OM
Huang et al., 2002	16330867	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16330868	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16330869	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16330919	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330937	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16330991	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16331025	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16331037	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16331081	*Synechocystis sp.*	CM	P
Fulda et al., 2000	16331081	*Synechocystis sp.*	P	P

			Localization	
Reference	NCBI GI	Organism	Laboratory	PSORTb
Huang et al., 2002	16331145	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16331157	*Synechocystis sp.*	CM	CM/P
Fulda et al., 2000	16331169	*Synechocystis sp.*	P	P
Fulda et al., 2000	16331198	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16331204	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16331258	*Synechocystis sp.*	P	C
Fulda et al., 2000	16331259	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16331342	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16331346	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16331360	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16331369	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16331395	*Synechocystis sp.*	CM	Unknown
Huang et al., 2002	16331445	*Synechocystis sp.*	CM	P
Fulda et al., 2000	16331473	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16331484	*Synechocystis sp.*	P	P
Huang et al., 2002	16331566	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16331612	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16331629	*Synechocystis sp.*	CM	OM
Fulda et al., 2000	16331677	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16331728	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16331793	*Synechocystis sp.*	CM	P
Fulda et al., 2000	16331793	*Synechocystis sp.*	P	P
Fulda et al., 2000	16331950	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16331998	*Synechocystis sp.*	CM	C
Huang et al., 2002	16332004	*Synechocystis sp.*	CM	Unknown
Fulda et al., 2000	16332232	*Synechocystis sp.*	P	Unknown
Fulda et al., 2000	16332251	*Synechocystis sp.*	P	Unknown
Huang et al., 2002	16332322	*Synechocystis sp.*	CM	Unknown
Dukan et al., 1998	17380384	*E. coli*	C	C
Nouwens et al., 2002	17865491	*P. aeruginosa*	OM	C
Molloy et al., 2001	20141233	*S. typhimurium*	OM	OM
Molloy et al., 2001	20141635	*S. typhimurium*	OM	OM
Molloy et al., 2001	20141670	*S. typhimurium*	OM	OM
Molloy et al., 2001	20141731	*S. typhimurium*	OM	OM
Nouwens et al., 2002	25008883	*P. aeruginosa*	OM	OM
Fulda et al., 2000	27805660	*Synechocystis sp.*	P	P
Molloy et al., 2000	32172423	*E. coli*	OM	OM

Appendix F

Percentage of proteins at each localization site for the 236 bacterial genomes analyzed using PSORTb v.2.0. Gram-negative organisms appear first, followed by Gram-positive organisms. Organisms are arranged alphabetically within their Gram grouping.

	Percentage of Proteome							
Organism (Proteome Size)	**C**	**CM**	**CW**	**P**	**OM**	**EC**	**Unknown**	**Mutiple**
A. sp. ADP1 (3325)	35.2	17.4	N/A	1.2	3.6	0.2	41.2	1.2
A. tumefaciens str. C58 (Cereon) (4554)	36.5	20.4	N/A	3.1	0.9	0.2	36.7	2.2
A. tumefaciens str. C58 (U. Washington) (4661)	36.9	19.7	N/A	3.1	0.9	0.2	36.9	2.2
A. variabilis ATCC 29413 (5039)	29.1	17.9	N/A	1.0	1.8	0.6	48.0	1.7
A. marginale str. St. Maries (949)	36.9	20.7	N/A	1.1	2.0	0.2	37.1	2.1
A. aeolicus VF5 (1529)	54.9	16.0	N/A	1.4	1.6	0.4	23.9	1.9
A. sp. EbN1 (4133)	38.4	14.9	N/A	1.6	1.3	0.3	41.8	1.9
B. fragilis NCTC 9434 (4189)	30.2	15.6	N/A	1.1	4.4	0.3	46.3	2.1
B. fragilis YCH46 (4578)	28.3	14.4	N/A	1.0	4.2	0.3	49.9	2.0
B. thetaiotaomicron VPI-5482 (4778)	28.4	14.7	N/A	1.0	5.2	0.5	48.2	2.0
B. henselae str. Houston-1 (1488)	35.0	15.9	N/A	1.3	2.0	0.1	43.9	1.8
B. quintana str. Toulouse (1142)	37.0	18.1	N/A	1.3	1.8	0.2	38.8	2.8
B. bacteriovorus HD100 (3587)	24.0	14.6	N/A	1.6	2.8	0.9	53.2	3.0
B. bronchiseptica RB50 (4994)	33.4	18.3	N/A	3.5	2.0	0.3	39.4	3.0
B. parapertussis 12822 (4185)	34.9	18.9	N/A	3.8	2.1	0.3	37.0	3.1
B. pertussis Tohama I (3436)	31.7	17.4	N/A	3.2	2.0	0.3	42.8	2.6
B. burgdorferi B31 (851)	30.4	20.1	N/A	1.2	5.4	0.5	40.8	1.6
B. garinii PBi (832)	30.3	19.7	N/A	1.0	3.4	0.5	43.8	1.4
B. japonicum USDA 110 (8317)	32.0	18.4	N/A	2.6	1.3	0.3	42.9	2.5
B. abortus biovar 1 str. 9-941 (3085)	36.6	17.2	N/A	2.6	1.0	0.1	40.2	2.5
B. melitensis 16M (3198)	37.2	18.1	N/A	2.7	1.0	0.1	38.5	2.5
B. suis 1330 (3264)	34.4	17.0	N/A	2.5	1.0	0.1	42.6	2.5
B. aphidicola str. APS (Acyrthosiphon pisum) (564)	34.9	14.9	N/A	0.9	2.0	0.2	46.1	1.1
B. aphidicola str. Bp (Baizongia pistaciae) (504)	30.0	17.5	N/A	1.0	1.6	0.2	48.2	1.6
B. aphidicola str. Sg (Schizaphis graminum) (546)	35.2	14.5	N/A	1.5	1.8	0.2	45.6	1.3
B. mallei ATCC 23344 (4764)	36.7	16.3	N/A	2.1	2.1	0.6	39.8	2.5
B. pseudomallei K96243 (5729)	33.9	18.3	N/A	2.3	2.6	0.7	39.7	2.7
C. jejuni RM1221 (1838)	31.1	15.7	N/A	1.6	3.0	0.3	46.7	1.6
C. jejuni subsp. jejuni NCTC 11168 (1634)	31.8	17.6	N/A	1.9	3.4	0.4	43.2	1.7
C. Blochmannia floridanus (583)	36.0	18.4	N/A	0.9	1.4	0.2	41.9	1.4
C. Blochmannia pennsylvanicus str. BPEN (610)	39.5	17.5	N/A	0.8	1.5	0.2	38.5	2.0
C. Pelagibacter ubique HTCC1062 (1354)	30.3	16.9	N/A	1.3	2.0	0.2	47.9	1.6
C. crescentus CB15 (3737)	33.2	15.6	N/A	1.9	2.8	0.4	43.8	2.3
C. muridarum Nigg (904)	33.6	16.3	N/A	0.7	2.5	0.1	43.8	3.0
C. trachomatis D/UW-3/CX (895)	35.4	16.0	N/A	1.0	2.5	0.2	43.0	1.9
C. abortus S26/3 (932)	34.6	18.0	N/A	0.6	2.3	0.2	41.6	2.7
C. caviae GPIC (998)	31.9	16.0	N/A	0.9	2.5	0.2	45.2	3.3
C. pneumoniae AR39 (1112)	30.6	15.3	N/A	0.8	2.4	0.1	48.3	2.5
C. pneumoniae CWL029 (1054)	32.1	16.2	N/A	0.9	2.6	0.1	45.5	2.8
C. pneumoniae J138 (1069)	31.7	16.0	N/A	0.8	2.6	0.1	45.2	3.6
C. pneumoniae TW-183 (1113)	30.6	15.4	N/A	0.8	2.4	0.1	47.8	3.0
C. tepidum TLS (2252)	38.0	14.3	N/A	0.9	1.2	0.2	44.0	1.3
C. violaceum ATCC 12472 (4407)	35.4	17.7	N/A	2.1	1.8	0.8	39.6	2.5
C. psychrerythraea 34H (4910)	26.9	15.6	N/A	1.6	3.5	0.5	50.4	1.5
C. burnetii RSA 493 (2009)	33.0	17.1	N/A	0.6	0.9	0.2	47.5	0.8
D. aromatica RCB (4171)	34.7	19.4	N/A	2.7	2.0	0.2	39.0	2.0
D. psychrophila LSv54 (3118)	36.3	18.9	N/A	1.5	1.4	0.3	39.6	2.0
D. vulgaris subsp. vulgaris str. Hildenborough (3379)	41.4	14.9	N/A	2.0	0.7	0.2	38.8	2.0
E. canis str. Jake (925)	29.1	18.2	N/A	0.5	3.1	0.3	46.9	1.8
E. ruminantium str. Gardel (950)	26.8	16.5	N/A	0.6	3.2	0.3	51.1	1.5
E. ruminantium str. Welgevonden (888)	28.4	16.8	N/A	0.7	3.0	0.2	49.1	1.8
E. ruminantium str. Welgevonden (958)	27.1	15.9	N/A	0.6	2.8	0.2	51.7	1.7

	Percentage of Proteome							
Organism (Proteome Size)	C	CM	CW	P	OM	EC	Unknown	Mutiple
E. carotovora subsp. atroseptica SCRI1043 (4472)	34.1	19.0	N/A	2.8	2.0	0.6	39.5	2.1
E. coli CFT073 (5379)	34.2	16.7	N/A	2.7	2.0	0.7	42.2	1.4
E. coli K12 (4311)	36.5	19.7	N/A	3.3	2.1	0.5	36.6	1.3
E. coli O157:H7 (5253)	34.6	16.6	N/A	2.8	2.2	0.6	41.6	1.5
E. coli O157:H7 EDL933 (5324)	33.9	16.8	N/A	2.8	2.4	0.6	42.1	1.5
F. tularensis subsp. tularensis Schu 4 (1603)	29.9	18.8	N/A	0.9	2.5	0.3	46.0	1.6
F. nucleatum subsp. nucleatum ATCC 25586 (2067)	39.3	18.0	N/A	1.0	3.8	0.1	34.8	3.0
G. sulfurreducens PCA (3445)	43.2	18.6	N/A	1.6	1.4	0.7	32.5	2.1
G. violaceus PCC 7421 (4430)	34.6	14.6	N/A	1.1	2.3	0.3	45.1	2.0
G. oxydans 621H (2432)	36.7	15.5	N/A	1.4	2.4	0.6	41.3	2.1
H. ducreyi 35000HP (1717)	33.8	14.6	N/A	1.9	2.2	0.2	46.0	1.4
H. influenzae 86-028NP (1791)	38.5	16.3	N/A	2.4	1.8	0.4	38.8	1.8
H. influenzae Rd KW20 (1657)	40.6	17.9	N/A	2.7	1.9	0.2	35.0	1.6
H. hepaticus ATCC 51449 (1875)	30.9	15.0	N/A	1.2	2.6	0.6	47.7	1.9
H. pylori 26695 (1576)	33.6	15.1	N/A	0.8	4.4	0.8	43.7	1.7
H. pylori J99 (1491)	33.8	16.0	N/A	0.9	4.2	0.9	42.2	1.9
I. loihiensis L2TR (2628)	35.7	17.4	N/A	1.5	3.1	0.4	39.5	2.4
L. pneumophila str. Lens (2878)	32.4	17.7	N/A	1.1	1.6	0.8	44.8	1.8
L. pneumophila str. Paris (3027)	32.2	18.0	N/A	1.1	1.6	0.7	44.7	1.6
L. pneumophila str. Philadelphia 1 (2942)	31.1	18.2	N/A	1.2	1.8	0.8	45.2	1.7
L. interrogans Copenhageni str. Fiocruz (3660)	27.7	16.2	N/A	0.9	2.8	0.7	50.3	1.5
L. interrogans Lai str. 56601 (4727)	23.8	12.9	N/A	0.6	2.4	0.6	58.6	1.2
M. succiniciproducens MBEL55E (2384)	33.1	16.5	N/A	3.0	1.8	0.2	44.0	1.4
M. florum L1 (683)	36.3	19.5	N/A	0.2	2.3	0.2	40.4	1.2
M. loti MAFF303099 (6746)	34.4	17.8	N/A	2.3	0.8	0.2	42.2	2.4
M. capsulatus str. Bath (2959)	41.2	17.3	N/A	2.0	1.4	0.4	35.0	2.7
M. gallisepticum R (726)	24.8	16.8	N/A	0.0	9.1	1.1	43.9	4.3
M. genitalium G-37 (484)	23.8	16.3	N/A	0.0	3.9	0.2	53.3	2.5
M. hyopneumoniae 232 (718)	18.8	24.2	N/A	0.7	7.7	0.1	45.5	2.9
M. hyopneumoniae 7448 (663)	21.7	21.3	N/A	0.6	8.1	0.2	45.6	2.6
M. hyopneumoniae J (665)	21.5	20.0	N/A	0.8	9.2	0.2	45.7	2.7
M. mobile 163K (633)	27.3	16.0	N/A	0.2	6.2	0.2	46.9	3.3
M. mycoides subsp. mycoides SC str. PG1 (1016)	22.7	15.2	N/A	0.0	7.0	0.1	52.8	2.3
M. penetrans HF-2 (1037)	23.8	15.9	N/A	0.1	10.1	0.1	43.9	6.1
M. pneumoniae M129 (689)	22.9	14.1	N/A	0.2	6.2	0.2	55.3	1.2
M. pulmonis UAB CTIP (782)	24.2	17.1	N/A	0.1	6.9	0.4	48.1	3.2
M. synoviae 53 (672)	21.0	13.7	N/A	0.2	11.0	0.2	50.9	3.1
N. gonorrhoeae FA 1090 (2002)	35.7	13.1	N/A	1.4	1.8	0.1	46.2	1.7
N. meningitidis MC58 (2079)	36.6	13.2	N/A	1.5	2.6	0.2	44.3	1.6
N. meningitidis Z2491 (2065)	36.3	13.3	N/A	1.6	2.1	0.2	45.0	1.6
N. winogradskyi Nb-255 (3122)	36.1	13.9	N/A	1.3	1.8	0.4	44.7	1.9
N. europaea ATCC 19718 (2461)	41.0	16.2	N/A	1.7	2.4	0.5	36.5	1.8
N. sp. PCC 7120 (5366)	28.7	16.9	N/A	1.1	1.9	0.5	49.0	1.9
O.Y. yellows phytoplasma OY-M (754)	23.9	15.8	N/A	0.4	1.1	0.1	57.6	1.2
P. sp. UWE25 (2031)	31.8	13.6	N/A	0.6	1.2	0.6	50.6	1.5
P. multocida subsp. multocida str. Pm70 (2015)	38.4	20.4	N/A	3.0	2.4	0.2	33.7	1.8
P. profundum SS9 (5413)	33.3	17.4	N/A	2.1	2.2	0.4	42.8	1.8
P. luminescens subsp. laumondii TTO1 (4683)	33.7	14.2	N/A	1.8	2.5	0.5	45.9	1.4
P. gingivalis W83 (1909)	38.4	13.5	N/A	0.6	2.9	0.4	42.9	1.5
P. marinus str. MIT 9313 (2265)	31.9	15.4	N/A	0.8	1.2	0.2	48.9	1.6
P. marinus str. NATL2A (1890)	31.4	13.0	N/A	0.7	1.6	0.4	51.5	1.4
P. marinus subsp. marinus str. CCMP1375 (1882)	31.4	13.3	N/A	0.6	1.4	0.2	51.6	1.5
P. marinus subsp. pastoris str. CCMP1986 (1712)	30.3	13.9	N/A	0.8	2.5	0.2	50.8	1.5
P. aeruginosa PAO1 (5567)	41.7	18.5	N/A	2.3	3.0	0.5	31.9	2.2
P. fluorescens Pf-5 (6137)	35.1	19.3	N/A	2.1	2.7	0.4	38.4	2.2
P. putida KT2440 (5350)	37.4	17.8	N/A	1.9	2.8	0.3	37.6	2.2
P. syringae pv. phaseolicola 1448A (4983)	34.7	18.3	N/A	2.2	2.3	0.4	39.7	2.3
P. syringae pv. syringae B728a (5090)	35.5	19.0	N/A	2.2	2.1	0.5	38.3	2.5
P. syringae pv. tomato str. DC3000 (5471)	34.9	17.3	N/A	2.0	2.2	0.4	40.8	2.6
P. arcticus 273-4 (2120)	31.1	16.2	N/A	1.4	1.9	0.2	48.3	0.9
R. eutropha JMP134 (5846)	34.1	17.8	N/A	3.2	2.1	0.3	40.3	2.3
R. solanacearum GMI1000 (3440)	33.7	15.1	N/A	2.1	1.7	0.4	44.5	2.4

	Percentage of Proteome							
Organism (Proteome Size)	C	CM	CW	P	OM	EC	Unknown	Mutiple
R. baltica SH 1 (7325)	26.4	11.3	N/A	0.6	0.8	0.3	57.9	2.8
R. palustris CGA009 (4814)	32.6	19.9	N/A	2.5	1.6	0.1	40.5	2.9
R. conorii str. Malish 7 (1374)	25.8	14.6	N/A	1.0	1.2	0.2	55.3	2.0
R. felis URRWXCal2 (1400)	30.1	16.3	N/A	1.0	1.9	0.2	48.6	1.9
R. prowazekii str. Madrid E (835)	29.9	20.8	N/A	1.1	2.3	0.1	43.8	1.9
R. typhi str. Wilmington (838)	30.4	19.1	N/A	1.0	1.8	0.1	45.8	1.8
S. enterica subsp. enterica Choleraesuis str. SC-B67 (4445)	33.9	18.2	N/A	2.8	1.7	0.7	41.1	1.6
S. enterica subsp. enterica Paratypi A ATC 9150 (4093)	35.5	19.1	N/A	3.1	1.7	0.6	38.2	1.8
S. enterica subsp. enterica Typhi str. CT18 (4395)	35.0	18.2	N/A	2.8	1.6	0.5	40.2	1.7
S. enterica subsp. enterica Typhi Ty2 (4318)	35.3	18.4	N/A	2.8	1.6	0.5	39.7	1.8
S. typhimurium LT2 (4425)	35.5	19.3	N/A	3.1	1.9	0.7	37.8	1.8
S. oneidensis MR-1 (4323)	30.5	16.1	N/A	2.4	2.6	0.5	45.8	2.1
S. flexneri 2a str. 2457T (4068)	35.1	17.1	N/A	2.9	1.5	0.3	41.6	1.5
S. flexneri 2a str. 301 (4180)	34.5	17.3	N/A	2.8	1.7	0.3	41.9	1.6
S. sonnei Ss046 (4223)	35.5	17.0	N/A	3.0	1.6	0.4	40.9	1.6
S. pomeroyi DSS-3 (3810)	39.1	17.5	N/A	2.3	0.8	0.5	37.2	2.7
S. meliloti 1021 (3341)	41.1	17.2	N/A	2.4	0.9	0.4	35.7	2.4
S. elongatus PCC 6301 (2525)	31.2	16.8	N/A	1.1	0.7	0.3	48.2	1.7
S. sp. WH 8102 (2517)	34.6	12.9	N/A	1.0	0.6	0.6	48.6	1.9
S. sp. PCC 6803 (3167)	32.0	17.1	N/A	1.4	1.3	0.3	46.4	1.7
T. elongatus BP-1 (2475)	33.5	17.6	N/A	0.9	0.7	0.1	45.7	1.7
T. maritima MSB8 (1858)	56.0	17.3	N/A	1.6	0.8	0.4	21.2	2.9
T. thermophilus HB27 (1982)	46.3	18.8	N/A	1.6	0.8	0.4	30.4	1.9
T. thermophilus HB8 (1973)	45.3	16.3	N/A	1.2	1.0	0.4	33.6	2.2
T. denitrificans ATCC 25259 (2827)	37.9	17.4	N/A	2.0	2.0	0.2	38.5	2.1
T. denticola ATCC 35405 (2767)	28.6	19.1	N/A	0.8	2.1	0.3	47.1	2.0
T. pallidum subsp. pallidum str. Nichols (1036)	39.4	17.1	N/A	1.5	1.7	0.3	38.1	1.9
U. parvum serovar 3 str. ATCC 700970 (614)	28.0	15.0	N/A	0.2	4.6	0.2	50.5	1.6
V. cholerae O1 biovar eltor str. N16961 (3835)	34.0	17.8	N/A	2.1	1.8	0.7	42.2	1.5
V. fischeri ES114 (3747)	35.7	19.2	N/A	2.3	2.9	0.5	37.7	1.8
V. parahaemolyticus RIMD 2210633 (4832)	35.2	17.4	N/A	2.3	2.7	0.8	39.7	2.0
V. vulnificus CMCP6 (4514)	36.0	18.4	N/A	2.2	2.5	0.8	38.2	2.0
V. vulnificus YJ016 (4955)	33.3	17.1	N/A	1.8	2.4	0.7	42.8	2.0
W. glossinidia endosymbiont of Glossina (611)	25.5	23.6	N/A	0.8	3.1	0.5	44.5	2.0
W. endosymbiont of Drosophila melanogaster (1195)	33.6	13.3	N/A	0.3	1.6	0.1	49.3	1.8
W. endosymbiont strain TRS of Brugia malayi (805)	36.0	16.2	N/A	0.5	2.0	0.1	43.2	2.0
W. succinogenes DSM 1740 (2044)	44.3	20.6	N/A	2.0	2.2	0.2	28.4	2.3
X. axonopodis pv. citri str. 306 (4312)	29.5	16.4	N/A	1.9	3.3	0.9	45.7	2.3
X. campestris pv. campestris str. 8004 (4273)	29.2	16.4	N/A	1.7	3.2	0.7	46.3	2.4
X. campestris pv. campestris str. ATCC 33913 (4181)	29.6	16.6	N/A	1.8	3.3	0.7	45.5	2.5
X. oryzae pv. oryzae KACC10331 (4637)	32.6	12.9	N/A	1.6	2.1	0.6	48.3	1.9
X. fastidiosa 9a5c (2766)	30.2	11.2	N/A	1.1	1.6	0.5	54.1	1.5
X. fastidiosa Temecula1 (2034)	34.9	14.1	N/A	1.4	2.1	0.6	44.7	2.2
Y. pestis biovar Medievalis str. 91001 (3895)	31.6	18.7	N/A	3.1	2.5	0.6	41.7	1.8
Y. pestis CO92 (3885)	32.4	18.4	N/A	3.1	2.7	0.6	41.1	1.7
Y. pestis KIM (4086)	30.5	17.8	N/A	3.0	2.7	0.6	43.8	1.6
Y. pseudotuberculosis IP 32953 (3901)	31.9	19.3	N/A	3.3	2.8	0.7	40.1	2.0
Z. mobilis subsp. mobilis ZM4 (1998)	29.6	13.3	N/A	1.2	2.8	0.9	50.3	2.1
B. anthracis str. A2012 (5544)	47.3	21.7	1.2	N/A	N/A	3.0	26.2	0.7
B. anthracis str. Ames (5311)	46.6	21.7	1.1	N/A	N/A	4.7	25.3	0.7
B. anthracis str. 'Ames Ancestor' (5309)	46.6	21.7	1.1	N/A	N/A	4.7	25.3	0.7
B. anthracis str. Sterne (5287)	48.6	23.2	1.2	N/A	N/A	2.1	24.1	0.8
B. cereus ATCC 10987 (5603)	45.1	21.3	1.0	N/A	N/A	6.5	25.1	0.9
B. cereus ATCC 14579 (5234)	48.5	21.7	0.9	N/A	N/A	3.5	24.4	1.0
B. cereus ZK (5134)	49.0	23.9	1.2	N/A	N/A	1.9	23.1	0.9
B. clausii KSM-K16 (4096)	54.4	23.1	0.3	N/A	N/A	1.7	18.1	2.5
B. halodurans C-125 (4066)	55.7	20.7	0.2	N/A	N/A	3.9	16.8	2.7
B. licheniformis ATCC 14580 (4196)	51.0	22.6	0.5	N/A	N/A	2.9	22.1	0.8
B. licheniformis ATCC 14580 (DSM 13) (4161)	50.8	22.4	0.5	N/A	N/A	3.3	22.1	0.9
B. subtilis subsp. subtilis str. 168 (4112)	50.6	22.2	0.5	N/A	N/A	2.6	23.5	0.7
B. thuringiensis serovar konkukian str. 97-27 (5117)	48.9	24.3	1.2	N/A	N/A	2.1	22.6	0.9
B. longum NCC2705 (1727)	54.7	20.5	1.3	N/A	N/A	1.1	21.8	0.6

	Percentage of Proteome							
Organism (Proteome Size)	**C**	**CM**	**CW**	**P**	**OM**	**EC**	**Unknown**	**Mutiple**
C. acetobutylicum ATCC 824 (3672)	46.9	21.1	0.8	N/A	N/A	2.5	28.3	0.4
C. perfringens str. 13 (2660)	54.5	22.1	0.8	N/A	N/A	2.1	19.1	1.3
C. tetani E88 (2373)	53.3	23.8	1.2	N/A	N/A	1.0	19.5	1.2
C. diphtheriae NCTC 13129 (2272)	52.3	19.0	0.8	N/A	N/A	2.2	24.6	1.2
C. efficiens YS-314 (2950)	54.9	18.4	0.4	N/A	N/A	1.9	23.0	1.4
C. glutamicum ATCC 13032 (2993)	52.1	20.1	0.5	N/A	N/A	1.6	24.4	1.4
C. glutamicum ATCC 13032 (3057)	50.9	19.9	0.5	N/A	N/A	2.4	24.9	1.5
C. jeikeium K411 (2137)	54.4	17.8	0.4	N/A	N/A	2.7	23.4	1.3
D. ethenogenes 195 (1580)	52.9	18.0	0.4	N/A	N/A	3.4	24.1	1.1
D. sp. CBDB1 (1458)	55.1	19.3	0.3	N/A	N/A	1.9	22.4	1.1
D. radiodurans R1 (2997)	49.3	15.2	0.6	N/A	N/A	1.8	32.7	0.5
E. faecalis V583 (3113)	48.7	20.0	1.5	N/A	N/A	3.7	25.8	0.3
G. kaustophilus HTA426 (3498)	57.0	19.8	0.3	N/A	N/A	3.0	18.8	1.0
L. acidophilus NCFM (1864)	44.4	21.0	2.2	N/A	N/A	3.2	28.9	0.3
L. johnsonii NCC 533 (1821)	44.0	24.5	1.4	N/A	N/A	1.0	28.9	0.2
L. plantarum WCFS1 (3009)	41.7	21.3	1.5	N/A	N/A	1.3	33.9	0.4
L. lactis subsp. lactis Il1403 (2321)	48.2	19.4	1.0	N/A	N/A	1.9	29.2	0.4
L. xyli subsp. xyli str. CTCB07 (2030)	48.6	18.1	0.3	N/A	N/A	2.9	29.6	0.5
L. innocua Clip11262 (2968)	54.8	19.7	1.6	N/A	N/A	1.4	21.9	0.6
L. monocytogenes EGD-e (2846)	55.1	21.1	1.9	N/A	N/A	1.5	19.7	0.7
L. monocytogenes str. 4b F2365 (2821)	53.4	21.2	2.0	N/A	N/A	2.7	20.0	0.8
M. avium subsp. paratuberculosis str. k10 (4350)	55.2	15.2	0.3	N/A	N/A	1.6	27.4	0.4
M. bovis AF2122/97 (3920)	52.7	15.1	0.2	N/A	N/A	3.5	28.0	0.5
M. leprae TN (1605)	53.2	15.6	0.0	N/A	N/A	1.9	28.7	0.7
M. tuberculosis CDC1551 (4187)	49.9	14.0	0.2	N/A	N/A	5.0	30.4	0.5
M. tuberculosis H37Rv (3927)	53.1	15.2	0.2	N/A	N/A	3.4	27.5	0.5
N. farcinica IFM 10152 (5683)	55.8	14.7	0.3	N/A	N/A	2.1	25.9	1.2
O. iheyensis HTE831 (3500)	50.1	23.2	0.3	N/A	N/A	2.3	22.0	2.2
P. acnes KPA171202 (2297)	54.0	21.1	0.3	N/A	N/A	1.7	22.6	0.4
S. aureus subsp. aureus COL (2615)	47.3	20.8	1.2	N/A	N/A	5.8	24.4	0.5
S. aureus subsp. aureus MRSA252 (2656)	48.4	20.9	1.0	N/A	N/A	4.8	24.5	0.5
S. aureus subsp. aureus MSSA476 (2579)	47.6	21.6	1.1	N/A	N/A	4.9	24.5	0.5
S. aureus subsp. aureus Mu50 (2714)	48.0	20.7	1.3	N/A	N/A	4.6	25.1	0.4
S. aureus subsp. aureus MW2 (2632)	47.3	21.3	1.2	N/A	N/A	5.2	24.5	0.5
S. aureus subsp. aureus N315 (2593)	48.1	21.3	1.2	N/A	N/A	4.5	24.5	0.4
S. epidermidis ATCC 12228 (2419)	46.7	20.3	1.2	N/A	N/A	7.3	23.9	0.5
S. epidermidis RP62A (2494)	48.0	18.8	1.2	N/A	N/A	6.6	24.9	0.5
S. haemolyticus JCSC1435 (2676)	49.3	19.9	0.9	N/A	N/A	4.3	25.2	0.3
S. saprophyticus subsp. saprophyticus ATCC 15305 (2446)	50.1	22.4	0.5	N/A	N/A	3.1	23.2	0.8
S. agalactiae 2603V/R (2124)	48.9	20.4	1.6	N/A	N/A	4.1	24.5	0.5
S. agalactiae NEM316 (2094)	49.8	21.8	2.1	N/A	N/A	1.6	24.3	0.5
S. mutans UA159 (1960)	48.2	21.4	0.7	N/A	N/A	3.4	26.2	0.2
S. pneumoniae R6 (2043)	52.4	20.5	0.9	N/A	N/A	2.6	23.0	0.6
S. pneumoniae TIGR4 (2094)	51.2	19.6	0.9	N/A	N/A	6.0	21.9	0.4
S. pyogenes M1 GAS (1697)	51.0	18.9	1.1	N/A	N/A	2.2	26.3	0.5
S. pyogenes MGAS10394 (1886)	49.3	17.4	1.1	N/A	N/A	3.0	29.0	0.2
S. pyogenes MGAS315 (1865)	50.3	17.0	1.2	N/A	N/A	2.0	29.2	0.3
S. pyogenes MGAS5005 (1865)	48.7	17.5	1.0	N/A	N/A	3.5	28.9	0.4
S. pyogenes MGAS6180 (1894)	47.7	17.6	1.4	N/A	N/A	3.2	29.8	0.3
S. pyogenes MGAS8232 (1845)	49.8	17.2	1.0	N/A	N/A	2.5	29.2	0.4
S. pyogenes SSI-1 (1861)	49.8	16.9	1.2	N/A	N/A	2.6	29.2	0.3
S. thermophilus CNRZ1066 (1915)	49.1	18.2	0.5	N/A	N/A	2.4	29.6	0.2
S. thermophilus LMG 18311 (1889)	49.3	18.7	0.7	N/A	N/A	2.2	28.9	0.2
S. avermitilis MA-4680 (7575)	52.5	16.7	0.4	N/A	N/A	2.9	26.9	0.6
S. coelicolor A3(2) (7769)	51.2	16.5	0.4	N/A	N/A	3.4	27.2	1.2
S. thermophilum IAM 14863 (3337)	57.0	20.5	0.4	N/A	N/A	2.4	18.0	1.7
T. tengcongensis MB4 (2588)	59.9	20.1	0.5	N/A	N/A	1.4	17.1	1.0
T. fusca YX (3110)	56.1	17.7	0.6	N/A	N/A	2.0	22.1	1.6
T. whipplei str. Twist (808)	52.4	21.5	0.6	N/A	N/A	1.4	23.4	0.8
T. whipplei TW08/27 (783)	53.1	22.6	0.8	N/A	N/A	1.7	20.7	1.1

REFERENCE LIST

Albertsson, P.A. 1956. Chromatography and partition of cells and cell fragments. *Nature* **177:** 771-774.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Anderson, D.M. and O. Schneewind. 1999. Type III machines of Gram-negative pathogens: injecting virulence factors into host cells and more. *Curr Opin Microbiol* **2:** 18-24.

Andrade, M.A., S.I. O'Donoghue, and B. Rost. 1998. Adaptation of protein surfaces to subcellular location. *J Mol Biol* **276:** 517-525.

Antelmann, H., H. Tjalsma, B. Voigt, S. Ohlmeier, S. Bron, J.M. van Dijl, and M. Hecker. 2001. A proteomic view on genome-based signal peptide predictions. *Genome Res* **11:** 1484-1502.

Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Bairoch, A. and B. Boeckmann. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* **19 Suppl:** 2247-2249.

Bairoch, A. and B. Boeckmann. 1992. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* **20 Suppl:** 2019-2022.

Bairoch, A. and B. Boeckmann. 1993. The SWISS-PROT protein sequence data bank, recent developments. *Nucleic Acids Res* **21:** 3093-3096.

Bairoch, A. and B. Boeckmann. 1994. The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res* **22:** 3578-3580.

Barns, S.M., C.F. Delwiche, J.D. Palmer, and N.R. Pace. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc Natl Acad Sci U S A* **93:** 9188-9193.

Bendtsen, J.D., T.T. Binnewies, P.F. Hallin, and D.W. Ussery. 2005. Genome update: prediction of membrane proteins in prokaryotic genomes. *Microbiology* **151:** 2119-2121.

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

Benson, D.A., I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, and D.L. Wheeler. 2002. GenBank. *Nucleic Acids Res* **30:** 17-20.

Berks, B.C. 1996. A common export pathway for proteins binding complex redox cofactors? *Mol Microbiol* **22:** 393-404.

Bernstein, H.D. 1998. Membrane protein biogenesis: the exception explains the rules. *Proc Natl Acad Sci U S A* **95:** 14587-14589.

Bernstein, H.D. 2000. The biogenesis and assembly of bacterial membrane proteins. *Curr Opin Microbiol* **3:** 203-209.

Bhasin, M., A. Garg, and G.P. Raghava. 2005. PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* **21:** 2522-2524.

Bina, J.E., F. Nano, and R.E. Hancock. 1997. Utilization of alkaline phosphatase fusions to identify secreted proteins, including potential efflux proteins and virulence factors from Helicobacter pylori. *FEMS Microbiol Lett* **148:** 63-68.

Blonder, J., M.B. Goshe, W. Xiao, D.G. Camp, 2nd, M. Wingerd, R.W. Davis, and R.D. Smith. 2004. Global analysis of the membrane subproteome of Pseudomonas aeruginosa using liquid chromatography-tandem mass spectrometry. *J Proteome Res* **3:** 434-444.

Boeckmann, B., A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31:** 365-370.

Boucher, Y., C.J. Douady, R.T. Papke, D.A. Walsh, M.E. Boudreau, C.L. Nesbo, R.J. Case, and W.F. Doolittle. 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annu Rev Genet* **37:** 283-328.

Bumann, D., S. Aksu, M. Wendland, K. Janek, U. Zimny-Arndt, N. Sabarth, T.F. Meyer, and P.R. Jungblut. 2002. Proteome analysis of secreted proteins of the gastric pathogen Helicobacter pylori. *Infect Immun* **70:** 3396-3403.

Burns, D.L. 2003. Type IV transporters of pathogenic bacteria. *Curr Opin Microbiol* **6:** 29-34.

Cachia, P.J. and R.S. Hodges. 2003. Synthetic peptide vaccine and antibody therapeutic development: prevention and treatment of Pseudomonas aeruginosa. *Biopolymers* **71:** 141-168.

Cai, Y.D., X.J. Liu, and K.C. Chou. 2002. Artificial neural network model for predicting protein subcellular location. *Comput Chem* **26:** 179-182.

Cai, Y.D., X.J. Liu, X.B. Xu, and K.C. Chou. 2000. Support vector machines for prediction of protein subcellular location. *Mol Cell Biol Res Commun* **4:** 230-233.

Cedano, J., P. Aloy, J.A. Perez-Pons, and E. Querol. 1997. Relation between amino acid composition and cellular location of proteins. *J Mol Biol* **266:** 594-600.

Chaddock, A.M., A. Mant, I. Karnauchov, S. Brink, R.G. Herrmann, R.B. Klosgen, and C. Robinson. 1995. A new type of signal peptide: central role of a twin-arginine motif in transfer signals for the delta pH-dependent thylakoidal protein translocase. *Embo J* **14:** 2715-2722.

Chakravarti, D.N., M.J. Fiske, L.D. Fletcher, and R.J. Zagursky. 2000. Mining genomes and mapping proteomes: identification and characterization of protein subunit vaccines. *Dev Biol (Basel)* **103:** 81-90.

Chalfie, M., Y. Tu, G. Euskirchen, W.W. Ward, and D.C. Prasher. 1994. Green fluorescent protein as a marker for gene expression. *Science* **263:** 802-805.

Chitlaru, T., N. Ariel, A. Zvi, M. Lion, B. Velan, A. Shafferman, and E. Elhanany. 2004. Identification of chromosomally encoded membranal polypeptides of Bacillus anthracis by a proteomic analysis: prevalence of proteins containing S-layer homology domains. *Proteomics* **4:** 677-691.

Chou, K.C. 2000. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophys Res Commun* **278:** 477-483.

Chou, K.C. and D.W. Elrod. 1998. Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem Biophys Res Commun* **252:** 63-68.

Chou, K.C. and D.W. Elrod. 1999. Protein subcellular location prediction. *Protein Eng* **12:** 107-118.

Christie, P.J. 2001. Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. *Mol Microbiol* **40:** 294-305.

Chung, Y.S., F. Breidt, and D. Dubnau. 1998. Cell surface localization and processing of the ComG proteins, required for DNA binding during transformation of Bacillus subtilis. *Mol Microbiol* **29:** 905-913.

Claros, M.G. and G. von Heijne. 1994. TopPred II: an improved software for membrane protein structure predictions. *Comput Appl Biosci* **10:** 685-686.

Cordwell, S.J., A.S. Nouwens, and B.J. Walsh. 2001. Comparative proteomics of bacterial pathogens. *Proteomics* **1:** 461-472.

Cserzo, M., E. Wallin, I. Simon, G. von Heijne, and A. Elofsson. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng* **10:** 673-676.

Darzins, A. 1994. Characterization of a Pseudomonas aeruginosa gene cluster involved in pilus biosynthesis and twitching motility: sequence similarity to the chemotaxis proteins of enterics and the gliding bacterium Myxococcus xanthus. *Mol Microbiol* **11:** 137-153.

de Cock, H., D. Hekstra, and J. Tommassen. 1990a. In vitro trimerization of outer membrane protein PhoE. *Biochimie* **72:** 177-182.

de Cock, H., R. Hendriks, T. de Vrije, and J. Tommassen. 1990b. Assembly of an in vitro synthesized Escherichia coli outer membrane porin into its stable trimeric configuration. *J Biol Chem* **265:** 4646-4651.

de Gier, J.W. and J. Luirink. 2001. Biogenesis of inner membrane proteins in Escherichia coli. *Mol Microbiol* **40:** 314-322.

de Gier, J.W., P.A. Scotti, A. Saaf, Q.A. Valent, A. Kuhn, J. Luirink, and G. von Heijne. 1998. Differential use of the signal recognition particle translocase targeting pathway for inner membrane protein assembly in Escherichia coli. *Proc Natl Acad Sci U S A* **95:** 14646-14651.

Deber, C.M., C. Wang, L.P. Liu, A.S. Prior, S. Agrawal, B.L. Muskat, and A.J. Cuticchia. 2001. TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci* **10:** 212-219.

Delepelaire, P. and C. Wandersman. 1990. Protein secretion in gram-negative bacteria. The extracellular metalloprotease B from Erwinia chrysanthemi contains a C-terminal secretion signal analogous to that of Escherichia coli alpha-hemolysin. *J Biol Chem* **265:** 17118-17125.

Dilks, K., R.W. Rose, E. Hartmann, and M. Pohlschroder. 2003. Prokaryotic utilization of the twin-arginine translocation pathway: a genomic survey. *J Bacteriol* **185:** 1478-1483.

Dougan, G., G. Dowd, and M. Kehoe. 1983. Organization of K88ac-encoded polypeptides in the Escherichia coli cell envelope: use of minicells and outer membrane protein mutants for studying assembly of pili. *J Bacteriol* **153:** 364-370.

Drew, D., L. Froderberg, L. Baars, and J.W. de Gier. 2003. Assembly and overexpression of membrane proteins in Escherichia coli. *Biochim Biophys Acta* **1610:** 3-10.

Dukan, S., E. Turlin, F. Biville, G. Bolbach, D. Touati, J.C. Tabet, and J.C. Blais. 1998. Coupling 2D SDS-PAGE with CNBr cleavage and MALDI-TOFMS: a strategy applied to the identification of proteins induced by a hypochlorous acid stress in Escherichia coli. *Anal Chem* **70:** 4433-4440.

Dunn, S.D., E. Kellner, and H. Lill. 2001. Specific heterodimer formation by the cytoplasmic domains of the b and b' subunits of cyanobacterial ATP synthase. *Biochemistry* **40:** 187-192.

Dunn, S.D., D.T. McLachlin, and M. Revington. 2000. The second stalk of Escherichia coli ATP synthase. *Biochim Biophys Acta* **1458:** 356-363.

Dutt, M.J. and K.H. Lee. 2000. Proteomic analysis. *Curr Opin Biotechnol* **11:** 176-179.

Eisenberg, D., R.M. Weiss, and T.C. Terwilliger. 1984. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci U S A* **81:** 140-144.

Eisenhaber, F. and P. Bork. 1998. Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol* **8:** 169-170.

Elias, D.A., M.E. Monroe, M.J. Marshall, M.F. Romine, A.S. Belieav, J.K. Fredrickson, G.A. Anderson, R.D. Smith, and M.S. Lipton. 2005. Global detection and characterization of hypothetical proteins in Shewanella oneidensis MR-1 using LC-MS based proteomics. *Proteomics* **5:** 3120-3130.

Ertl, P., M. Wagner, E. Corton, and S.R. Mikkelsen. 2003. Rapid identification of viable Escherichia coli subspecies with an electrochemical screen-printed biosensor array. *Biosens Bioelectron* **18:** 907-916.

Fekkes, P. and A.J. Driessen. 1999. Protein targeting to the bacterial cytoplasmic membrane. *Microbiol Mol Biol Rev* **63:** 161-173.

Ferguson, A.D., E. Hofmann, J.W. Coulton, K. Diederichs, and W. Welte. 1998. Siderophore-mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide. *Science* **282:** 2215-2220.

Fernandez, R.C. and A.A. Weiss. 1994. Cloning and sequencing of a Bordetella pertussis serum resistance locus. *Infect Immun* **62:** 4727-4738.

Filloux, A., G. Michel, and M. Bally. 1998. GSP-dependent protein secretion in gram-negative bacteria: the Xcp system of Pseudomonas aeruginosa. *FEMS Microbiol Rev* **22:** 177-198.

Fischer, W., R. Haas, and S. Odenbreit. 2002. Type IV secretion systems in pathogenic bacteria. *Int J Med Microbiol* **292:** 159-168.

Fischetti, V.A. 2000. *Gram-positive pathogens*. ASM Press, Washington, D.C.

Folz, R.J. and J.I. Gordon. 1987. Computer-assisted predictions of signal peptidase processing sites. *Biochem Biophys Res Commun* **146:** 870-877.

Fridkin, S.K. and R.P. Gaynes. 1999. Antimicrobial resistance in intensive care units. *Clin Chest Med* **20:** 303-316, viii.

Fulda, S., F. Huang, F. Nilsson, M. Hagemann, and B. Norling. 2000. Proteomics of Synechocystis sp. strain PCC 6803. Identification of periplasmic proteins in cells grown at low and high salt concentrations. *Eur J Biochem* **267:** 5900-5907.

Gardy, J.L., M.R. Laird, F. Chen, S. Rey, C.J. Walsh, M. Ester, and F.S. Brinkman. 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21:** 617-623.

Gardy, J.L., C. Spencer, K. Wang, M. Ester, G.E. Tusnady, I. Simon, S. Hua, K. deFays, C. Lambert, K. Nakai, and F.S. Brinkman. 2003. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res* **31:** 3613-3617.

Ghigo, J.M. and C. Wandersman. 1994. A carboxyl-terminal four-amino acid motif is required for secretion of the metalloprotease PrtG through the Erwinia chrysanthemi protease secretion pathway. *J Biol Chem* **269:** 8979-8985.

Goshe, M.B., J. Blonder, and R.D. Smith. 2003. Affinity labeling of highly hydrophobic integral membrane proteins for proteome-wide analysis. *J Proteome Res* **2:** 153-161.

Govan, J.R. and V. Deretic. 1996. Microbial pathogenesis in cystic fibrosis: mucoid Pseudomonas aeruginosa and Burkholderia cepacia. *Microbiol Rev* **60:** 539-574.

Govorun, V.M. and A.I. Archakov. 2002. Proteomic technologies in modern biomedical science. *Biochemistry (Mosc)* **67:** 1109-1123.

Granseth, E., D.O. Daley, M. Rapp, K. Melen, and G. von Heijne. 2005. Experimentally constrained topology models for 51,208 bacterial inner membrane proteins. *J Mol Biol* **352:** 489-494.

Guina, T., S.O. Purvine, E.C. Yi, J. Eng, D.R. Goodlett, R. Aebersold, and S.I. Miller. 2003. Quantitative proteomic analysis indicates increased synthesis of a quinolone by Pseudomonas aeruginosa isolates from cystic fibrosis airways. *Proc Natl Acad Sci U S A* **100:** 2771-2776.

Hancock, R.E. and D.P. Speert. 2000. Antibiotic resistance in Pseudomonas aeruginosa: mechanisms and impact on treatment. *Drug Resist Updat* **3:** 247-255.

Hefty, P.S., S.E. Jolliff, M.J. Caimano, S.K. Wikel, and D.R. Akins. 2002. Changes in temporal and spatial patterns of outer surface lipoprotein expression generate population heterogeneity and antigenic diversity in the Lyme disease spirochete, Borrelia burgdorferi. *Infect Immun* **70:** 3468-3478.

Henderson, I.R., R. Cappello, and J.P. Nataro. 2000. Autotransporter proteins, evolution and redefining protein secretion. *Trends Microbiol* **8:** 529-532.

Henderson, I.R., F. Navarro-Garcia, and J.P. Nataro. 1998. The great escape: structure and function of the autotransporter proteins. *Trends Microbiol* **6:** 370-378.

Hermann, T., W. Pfefferle, C. Baumann, E. Busker, S. Schaffer, M. Bott, H. Sahm, N. Dusch, J. Kalinowski, A. Puhler, A.K. Bendt, R. Kramer, and A. Burkovski. 2001. Proteome analysis of Corynebacterium glutamicum. *Electrophoresis* **22:** 1712-1723.

Hinsa, S.M., M. Espinosa-Urgel, J.L. Ramos, and G.A. O'Toole. 2003. Transition from reversible to irreversible attachment during biofilm formation by Pseudomonas fluorescens WCS365 requires an ABC transporter and a large secreted protein. *Mol Microbiol* **49:** 905-918.

Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res* **27:** 215-219.

Hofmann, K. and W. Stoffel. 1992. PROFILEGRAPH: an interactive graphical tool for protein sequence analysis. *Comput Appl Biosci* **8:** 331-337.

Holland, I.B., L. Schmitt, and J. Young. 2005. Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway (review). *Mol Membr Biol* **22:** 29-39.

Hu, Y. and R.F. Murphy. 2004. Automated interpretation of subcellular patterns from immunofluorescence microscopy. *J Immunol Methods* **290:** 93-105.

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

Huang, F., I. Parmryd, F. Nilsson, A.L. Persson, H.B. Pakrasi, B. Andersson, and B. Norling. 2002. Proteomics of Synechocystis sp. strain PCC 6803: identification of plasma membrane proteins. *Mol Cell Proteomics* **1:** 956-966.

Hueck, C.J. 1998. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol Rev* **62:** 379-433.

Hui, L. 1992. Color set size problem with applications to string matching, combinatorial string matching. *Lecture Notes in Computer Science* **644:** 230-243.

Hulo, N., C.J. Sigrist, V. Le Saux, P.S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, and A. Bairoch. 2004. Recent improvements to the PROSITE database. *Nucleic Acids Res* **32:** D134-137.

Hultgren, S.J., F. Lindberg, G. Magnusson, J. Kihlberg, J.M. Tennent, and S. Normark. 1989. The PapG adhesin of uropathogenic Escherichia coli contains separate regions for receptor binding and for the incorporation into the pilus. *Proc Natl Acad Sci U S A* **86:** 4357-4361.

Joachims, T. 2002. SVMLight. http://svmlight.joachims.org

Jones, D.T., W.R. Taylor, and J.M. Thornton. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* **33:** 3038-3049.

Jonsson, A.P. 2001. Mass spectrometry for protein and peptide characterisation. *Cell Mol Life Sci* **58:** 868-884.

Juncker, A.S., H. Willenbrock, G. Von Heijne, S. Brunak, H. Nielsen, and A. Krogh. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12:** 1652-1662.

Juretic, D., L. Zoranic, and D. Zucic. 2002. Basic charge clusters and predictions of membrane protein topology. *J Chem Inf Comput Sci* **42:** 620-632.

Kall, L., A. Krogh, and E.L. Sonnhammer. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338:** 1027-1036.

Kawabe, T., E. Fujihira, and A. Yamaguchi. 2000. Molecular construction of a multidrug exporter system, AcrAB: molecular interaction between AcrA and AcrB, and cleavage of the N-terminal signal sequence of AcrA. *J Biochem (Tokyo)* **128:** 195-200.

Klauser, T., J. Pohlner, and T.F. Meyer. 1990. Extracellular transport of cholera toxin B subunit using Neisseria IgA protease beta-domain: conformation-dependent outer membrane translocation. *Embo J* **9:** 1991-1999.

Klauser, T., J. Pohlner, and T.F. Meyer. 1992. Selective extracellular release of cholera toxin B subunit by Escherichia coli: dissection of Neisseria Iga beta-mediated outer membrane transport. *Embo J* **11:** 2327-2335.

Kleinschmidt, J.H. and L.K. Tamm. 2002. Secondary and tertiary structure formation of the beta-barrel membrane protein OmpA is synchronized and depends on membrane thickness. *J Mol Biol* **324:** 319-330.

Klenk, H.P., M. Spitzer, T. Ochsenreiter, and G. Fuellen. 2004. Phylogenomics of hyperthermophilic Archaea and Bacteria. *Biochem Soc Trans* **32:** 175-178.

Knapp, J.E., D. Carroll, J.E. Lawson, S.R. Ernst, L.J. Reed, and M.L. Hackert. 2000. Expression, purification, and structural analysis of the trimeric form of the catalytic domain of the Escherichia coli dihydrolipoamide succinyltransferase. *Protein Sci* **9:** 37-48.

Knapp, J.E., D.T. Mitchell, M.A. Yazdi, S.R. Ernst, L.J. Reed, and M.L. Hackert. 1998. Crystal structure of the truncated cubic core component of the Escherichia coli 2-oxoglutarate dehydrogenase multienzyme complex. *J Mol Biol* **280:** 655-668.

Koronakis, V. and C. Hughes. 1993. Bacterial signal peptide-independent protein export: HlyB-directed secretion of hemolysin. *Semin Cell Biol* **4:** 7-15.

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

Kumar, R.B., Y.H. Xie, and A. Das. 2000. Subcellular localization of the Agrobacterium tumefaciens T-DNA transport pore proteins: VirB8 is essential for the assembly of the transport pore. *Mol Microbiol* **36:** 608-617.

Kyte, J. and R.F. Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157:** 105-132.

Ladunga, I., F. Czako, I. Csabai, and T. Geszti. 1991. Improving signal peptide prediction accuracy by simulated neural network. *Comput Appl Biosci* **7:** 485-487.

Landau, G. and U. Vishkin. 1989. Fast parallel and serial approximate string matching, *J Algorithms*, **10**: 157-169.

LaPointe, C.F. and R.K. Taylor. 2000. The type 4 prepilin peptidases comprise a novel family of aspartic acid proteases. *J Biol Chem* **275:** 1502-1510.

Lawrence, J.G. and J.R. Roth. 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143:** 1843-1860.

Lay, J.O., Jr. 2001. MALDI-TOF mass spectrometry of bacteria. *Mass Spectrom Rev* **20:** 172-194.

Letoffe, S., J.M. Ghigo, and C. Wandersman. 1994. Secretion of the Serratia marcescens HasA protein by an ABC transporter. *J Bacteriol* **176:** 5372-5377.

Letoffe, S. and C. Wandersman. 1992. Secretion of CyaA-PrtB and HlyA-PrtB fusion proteins in Escherichia coli: involvement of the glycine-rich repeat domain of Erwinia chrysanthemi protease B. *J Bacteriol* **174:** 4920-4927.

Lewenza, S., J.L. Gardy, F.S. Brinkman, and R.E. Hancock. 2005. Genome-wide identification of Pseudomonas aeruginosa exported proteins using a consensus computational strategy combined with a laboratory-based PhoA fusion screen. *Genome Res* **15:** 321-329.

Light, S., P. Kraulis, and A. Elofsson. 2005. Preferential attachment in the evolution of metabolic networks. *BMC Genomics* **6:** 159.

Lin, C. 2003. LibSVM. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Lloyd, S.A., M. Norman, R. Rosqvist, and H. Wolf-Watz. 2001. Yersinia YopE is targeted for type III secretion by N-terminal, not mRNA, signals. *Mol Microbiol* **39:** 520-531.

Locher, K.P., B. Rees, R. Koebnik, A. Mitschler, L. Moulinier, J.P. Rosenbusch, and D. Moras. 1998. Transmembrane signaling across the ligand-gated FhuA receptor: crystal structures of free and ferrichrome-bound states reveal allosteric changes. *Cell* **95:** 771-778.

Lory, S. 1994. Leader peptidases of type IV prepilins and related proteins. In G. von Heijne, *Signal peptidases,* pp. 31-44. R.G. Landes Co., Austin, Tex.

Lu, P., D. Szafron, R. Greiner, D.S. Wishart, A. Fyshe, B. Pearcy, B. Poulin, R. Eisner, D. Ngo, and N. Lamb. 2005. PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. *Nucleic Acids Res* **33:** D147-153.

Lu, Z., D. Szafron, R. Greiner, P. Lu, D.S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner. 2004. Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* **20:** 547-556.

Luirink, J., S. High, H. Wood, A. Giner, D. Tollervey, and B. Dobberstein. 1992. Signal-sequence recognition by an Escherichia coli ribonucleoprotein complex. *Nature* **359:** 741-743.

Luirink, J., C.M. ten Hagen-Jongman, C.C. van der Weijden, B. Oudega, S. High, B. Dobberstein, and R. Kusters. 1994. An alternative protein targeting pathway in Escherichia coli: studies on the role of FtsY. *Embo J* **13:** 2289-2296.

Lukashin, A.V. and M. Borodovsky. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26:** 1107-1115.

Ma, Q., Y. Zhai, J.C. Schneider, T.M. Ramseier, and M.H. Saier, Jr. 2003. Protein secretion systems of Pseudomonas aeruginosa and P fluorescens. *Biochim Biophys Acta* **1611:** 223-233.

Mackman, N., K. Baker, L. Gray, R. Haigh, J.M. Nicaud, and I.B. Holland. 1987. Release of a chimeric protein into the medium from Escherichia coli using the C-terminal secretion signal of haemolysin. *Embo J* **6:** 2835-2841.

Manoil, C. and J. Beckwith. 1985. TnphoA: a transposon probe for protein export signals. *Proc Natl Acad Sci U S A* **82:** 8129-8133.

Manoil, C. and B. Traxler. 1995. Membrane protein assembly: genetic, evolutionary and medical perspectives. *Annu Rev Genet* **29:** 131-150.

Marques, M.A., B.J. Espinosa, E.K. Xavier da Silveira, M.C. Pessolani, A. Chapeaurouge, J. Perales, K.M. Dobos, J.T. Belisle, J.S. Spencer, and P.J. Brennan. 2004. Continued proteomic analysis of Mycobacterium leprae subcellular fractions. *Proteomics* **4:** 2942-2953.

Mattick, J.S. 2002. Type IV pili and twitching motility. *Annu Rev Microbiol* **56:** 289-314.

McGeoch, D.J. 1985. On the predictive recognition of signal peptide sequences. *Virus Res* **3:** 271-286.

Mdluli, K.E., J.D. Treit, V.J. Kerr, and F.E. Nano. 1995. New vectors for the in vitro generation of alkaline phosphatase fusions to proteins encoded by G+C-rich DNA. *Gene* **155:** 133-134.

Menne, K.M., H. Hermjakob, and R. Apweiler. 2000. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* **16:** 741-742.

Milburn, M.V., G.G. Prive, D.L. Milligan, W.G. Scott, J. Yeh, J. Jancarik, D.E. Koshland, Jr., and S.H. Kim. 1991. Three-dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand. *Science* **254:** 1342-1347.

Molloy, M.P., B.R. Herbert, M.B. Slade, T. Rabilloud, A.S. Nouwens, K.L. Williams, and A.A. Gooley. 2000. Proteomic analysis of the Escherichia coli outer membrane. *Eur J Biochem* **267:** 2871-2881.

Molloy, M.P., N.D. Phadke, J.R. Maddock, and P.C. Andrews. 2001. Two-dimensional electrophoresis and peptide mass fingerprinting of bacterial outer membrane proteins. *Electrophoresis* **22:** 1686-1696.

Morse, S.A. 1978. The biology of the gonococcus. *CRC Crit Rev Microbiol* **7:** 93-189.

Muller, M. and R.B. Klosgen. 2005. The Tat pathway in bacteria and chloroplasts (review). *Mol Membr Biol* **22:** 113-121.

Murakami, Y., M. Imai, H. Nakamura, and F. Yoshimura. 2002. Separation of the outer membrane and identification of major outer membrane proteins from Porphyromonas gingivalis. *Eur J Oral Sci* **110:** 157-162.

Nair, R. and B. Rost. 2002a. Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics* **18 Suppl 1:** S78-86.

Nair, R. and B. Rost. 2002b. Sequence conserved for subcellular localization. *Protein Sci* **11:** 2836-2847.

Nair, R. and B. Rost. 2005. Mimicking cellular sorting improves prediction of subcellular localization. *J Mol Biol* **348:** 85-100.

Nakai, K. 2000. Protein sorting signals and prediction of subcellular localization. *Adv Protein Chem* **54:** 277-344.

Nakai, K. and P. Horton. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24:** 34-36.

Nakai, K. and M. Kanehisa. 1991. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins* **11:** 95-110.

Nakai, K. and M. Kanehisa. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* **14:** 897-911.

Neidhardt, F.C. and R. Curtiss. 1996. *Escherichia coli and Salmonella: cellular and molecular biology.* ASM Press, Washington, D.C.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne. 1997. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst* **8:** 581-599.

Nielsen, H. and A. Krogh. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol* **6:** 122-130.

Nishiyama, K., M. Hanada, and H. Tokuda. 1994. Disruption of the gene encoding p12 (SecG) reveals the direct involvement and important function of SecG in the protein translocation of Escherichia coli at low temperature. *Embo J* **13:** 3272-3277.

Nouwens, A.S., S.J. Cordwell, M.R. Larsen, M.P. Molloy, M. Gillings, M.D. Willcox, and B.J. Walsh. 2000. Complementing genomics with proteomics: the membrane subproteome of Pseudomonas aeruginosa PAO1. *Electrophoresis* **21:** 3797-3809.

Nouwens, A.S., M.D. Willcox, B.J. Walsh, and S.J. Cordwell. 2002. Proteomic comparison of membrane and extracellular proteins from invasive (PAO1) and cytotoxic (6206) strains of Pseudomonas aeruginosa. *Proteomics* **2:** 1325-1346.

Ochsner, U.A., A. Snyder, A.I. Vasil, and M.L. Vasil. 2002. Effects of the twin-arginine translocase on secretion of virulence factors, stress response, and pathogenesis. *Proc Natl Acad Sci U S A* **99:** 8312-8317.

Pal, C., B. Papp, and M.J. Lercher. 2005a. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* **37:** 1372-1375.

Pal, C., B. Papp, and M.J. Lercher. 2005b. Horizontal gene transfer depends on gene content of the host. *Bioinformatics* **21 Suppl 2:** ii222-ii223.

Pasquali, C., I. Fialka, and L.A. Huber. 1999. Subcellular fractionation, electromigration analysis and mapping of organelles. *J Chromatogr B Biomed Sci Appl* **722:** 89-102.

Peng, J. and S.P. Gygi. 2001. Proteomics: the move to mixtures. *J Mass Spectrom* **36:** 1083-1091.

Persson, B. and P. Argos. 1997. Prediction of membrane protein topology utilizing multiple sequence alignments. *J Protein Chem* **16:** 453-457.

Pohlner, J., R. Halter, and T.F. Meyer. 1987. Neisseria gonorrhoeae IgA protease. Secretion and implications for pathogenesis. *Antonie Van Leeuwenhoek* **53:** 479-484.

Poole, K., K. Krebes, C. McNally, and S. Neshat. 1993. Multiple antibiotic resistance in Pseudomonas aeruginosa: evidence for involvement of an efflux operon. *J Bacteriol* **175:** 7363-7372.

Popowicz, A.M. and P.F. Dash. 1988. SIGSEQ: a computer program for predicting signal sequence cleavage sites. *Comput Appl Biosci* **4:** 405-406.

Pugsley, A.P. 1993a. The complete general secretory pathway in gram-negative bacteria. *Microbiol Rev* **57:** 50-108.

Pugsley, A.P. 1993b. Processing and methylation of PulG, a pilin-like component of the general secretory pathway of Klebsiella oxytoca. *Mol Microbiol* **9:** 295-308.

Quentin, Y., G. Fichant, and F. Denizot. 1999. Inventory, assembly and analysis of Bacillus subtilis ABC transport systems. *J Mol Biol* **287:** 467-484.

Reinhardt, A. and T. Hubbard. 1998. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res* **26:** 2230-2236.

Rey, S., M. Acab, J.L. Gardy, M.R. Laird, K. deFays, C. Lambert, and F.S. Brinkman. 2005a. PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Res* **33:** D164-168.

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

Rost, B., P. Fariselli, and R. Casadio. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci* **5:** 1704-1718.

Saleh, M.T., M. Fillon, P.J. Brennan, and J.T. Belisle. 2001. Identification of putative exported/secreted proteins in prokaryotic proteomes. *Gene* **269:** 195-204.

Sandkvist, M. 2001. Type II secretion and pathogenesis. *Infect Immun* **69:** 3523-3535.

Santoni, V., M. Molloy, and T. Rabilloud. 2000. Membrane proteins and proteomics: un amour impossible? *Electrophoresis* **21:** 1054-1070.

Schaffer, S., B. Weil, V.D. Nguyen, G. Dongmann, K. Gunther, M. Nickolaus, T. Hermann, and M. Bott. 2001. A high-resolution reference map for cytoplasmic and membrane-associated proteins of Corynebacterium glutamicum. *Electrophoresis* **22:** 4404-4422.

Schatz, G. and B. Dobberstein. 1996. Common principles of protein translocation across membranes. *Science* **271:** 1519-1526.

Schleiff, E. and J. Soll. 2005. Membrane protein insertion: mixing eukaryotic and prokaryotic concepts. *EMBO Rep* **6:** 1023-1027.

Schneider, G. 1999. How many potentially secreted proteins are contained in a bacterial genome? *Gene* **237:** 113-121.

Schneider, G., S. Rohlk, and P. Wrede. 1993. Analysis of cleavage-site patterns in protein precursor sequences with a perceptron-type neural network. *Biochem Biophys Res Commun* **194:** 951-959.

Schneider, G. and P. Wrede. 1993. Development of artificial neural filters for pattern recognition in protein sequences. *J Mol Evol* **36:** 586-595.

Schulz, G.E. 2002. The structure of bacterial outer membrane proteins. *Biochim Biophys Acta* **1565:** 308-317.

She, R., F. Chen, K. Wang, M. Ester, J.L. Gardy, and F.S.L. Brinkman. 2003. Frequent subsequence-based prediction of outer membrane proteins. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 436-445. ACM Press, New York, NY.

Shi, S.Y., X.H. Cai, and D.F. Ding. 2005. Identification and categorization of horizontally transferred genes in prokaryotic genomes. *Acta Biochim Biophys Sin (Shanghai)* **37:** 561-566.

Sinha, S., S. Arora, K. Kosalai, A. Namane, A.S. Pym and S.T. Cole. 2002. Proteome analysis of the plasma membrane of Mycobacterium tuberculosis. *Comp Funct Genom* **3:** 470-483.

Sonenshein, A.L., J.A. Hoch, and R. Losick. 2002. *Bacillus subtilis and its closest relatives: from genes to cells*. ASM Press, Washington, D.C.

Sonnhammer, E.L., G. von Heijne, and A. Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* **6:** 175-182.

Stasyk, T. and L.A. Huber. 2004. Zooming in: fractionation strategies in proteomics. *Proteomics* **4:** 3704-3716.

Stover, C.K., X.Q. Pham, A.L. Erwin, S.D. Mizoguchi, P. Warrener, M.J. Hickey, F.S. Brinkman, W.O. Hufnagle, D.J. Kowalik, M. Lagrou, R.L. Garber, L. Goltry, E. Tolentino, S. Westbrock-Wadman, Y. Yuan, L.L. Brody, S.N. Coulter, K.R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G.K. Wong, Z. Wu, I.T. Paulsen, J. Reizer, M.H. Saier, R.E. Hancock, S. Lory, and M.V. Olson. 2000. Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. *Nature* **406:** 959-964.

Strachan, T. and A.P. Read. 2004. *Human molecular genetics 3.* Garland Press, London; New York.

Takeyasu, K., H. Omote, S. Nettikadan, F. Tokumasu, A. Iwamoto-Kihara, and M. Futai. 1996. Molecular imaging of Escherichia coli F0F1-ATPase in reconstituted membranes using atomic force microscopy. *FEBS Lett* **392:** 110-113.

Tusnady, G.E. and I. Simon. 1998. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol* **283:** 489-506.

Tusnady, G.E. and I. Simon. 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17:** 849-850.

Vapnik, V.N. 2000. *The nature of statistical learning theory.* Springer-Verlag, New York.

Vogt, J. and G.E. Schulz. 1999. The structure of the outer membrane protein OmpX from Escherichia coli reveals possible mechanisms of virulence. *Structure* **7:** 1301-1309.

von Heijne, G. 1985. Signal sequences. The limits of variation. *J Mol Biol* **184:** 99-105.

von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res* **14:** 4683-4690.

von Heijne, G. 1988. Transcending the impenetrable: how proteins come to terms with membranes. *Biochim Biophys Acta* **947:** 307-333.

von Heijne, G. 1989. The structure of signal peptides from bacterial lipoproteins. *Protein Eng* **2:** 531-534.

von Heijne, G. 1992. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225:** 487-494.

von Heijne, G. 1994. Signals for protein targeting into and across membranes. *Subcell Biochem* **22:** 1-19.

Voulhoux, R., G. Ball, B. Ize, M.L. Vasil, A. Lazdunski, L.F. Wu, and A. Filloux. 2001. Involvement of the twin-arginine translocation system in protein secretion via the type II pathway. *Embo J* **20:** 6735-6741.

Wang, J., G. Chirn, T. Marr, B. Shapiro, D. Shasha and K. Zhang. 1994. Combinatorial pattern discovery for scientific data: some preliminary results. SIGMOD-94, Minnesota, USA.

Wang, J., W.K. Sung, A. Krishnan, and K.B. Li. 2005. Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines. *BMC Bioinformatics* **6:** 174.

Wheeler, D.L., C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler, T.A. Tatusova, and B.A. Rapp. 2000. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **28:** 10-14.

White, S.H. and W.C. Wimley. 1999. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct* **28:** 319-365.

Yamaguchi, K., F. Yu, and M. Inouye. 1988. A single amino acid determinant of the membrane localization of lipoproteins in E. coli. *Cell* **53:** 423-432.

Yu, C.S., C.J. Lin, and J.K. Hwang. 2004. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci* **13:** 1402-1406.

Zgurskaya, H.I. and H. Nikaido. 2000. Cross-linked complex between oligomeric periplasmic lipoprotein AcrA and the inner-membrane-associated multidrug efflux pump AcrB from Escherichia coli. *J Bacteriol* **182:** 4264-4267.

Ziebandt, A.K., H. Weber, J. Rudolph, R. Schmid, D. Hoper, S. Engelmann, and M. Hecker. 2001. Extracellular proteins of Staphylococcus aureus and the role of SarA and sigma B. *Proteomics* **1:** 480-493.