Identification of Surface Structures Common to Gram-Negative Bacteria that are Suitable for Vaccine Development

Dissertation

der Mathematish-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Nagarajan Paramasivam
aus Madurai, India

Tübingen

2012

Tag der mündlichen Qualifikation:	09.10.2012
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	PD. Dr. Dirk Linke
2. Berichterstatter:	Prof. Dr. Friedrich Götz

To mom and dad

(To amma and appa)

Acknowledgements

I am very grateful to my supervisor Dr. Dirk Linke, whose encouragement, support and guidance helped me to complete my PhD.

I am very thankful to Prof. Dr. Andrei Lupas for providing me the opportunity to work in the department. It was a great privilege to work in the department and in the Max Planck Institute.

I am pleased to have Prof. Dr. Friedrich Götz, Prof. Dr. Volkmar Braun and Prof. Dr. Boris Maček as my thesis committee members.

I am very happy to work with Macrus Thein, Thomas Arnold, Iwan Grin and Jack Leo, they are very wonderful colleagues, whose company, help and their patience to read my manuscripts provided an great time through my PhD.

I am very thankful for Vikram Alva and Stanislaw Dunin-Horkawicz of their suggestions and discussions regarding the work. I am also thankful to Vikram for reading my manuscripts and providing valuable suggestions.

And Dr. Michael Habeck's suggestions and help in analyzing were really helpful in concluding the analysis of the C-terminal insertion signals.

It was really nice to have such a great and wonderful friends in Tübingen their company helped me to overcome the difficult times and made good time more memorable.

The company of the table football team and all the after-lunch-games we played were really great stress busters.

And my parents Gnanavalli and Paramasivam, whose dedication, love and sacrifices brought me here and I dedicate this thesis to them. I am very thankful to my Uncle and Aunt who supported through my education.

For me to reach this point in my life a lot of people helped me and provided the necessary motivations, it will be a long list to name them all, but I will always remain thankful to you people.

Table of Contents

I. Abbreviations

IM – Inner membrane

OM – Outer membrane

OMP – Outer membrane protein

SCL – Subcellular localization

SCF – Subcellular fractionation

SP – Signal peptide

HLA – Human leukocyte antigen

MHC – Major histocompatibility complex

PMN – Polymorphonuclear leukocyte

PRR – Pattern recognition receptor

CD – Cluster of differentiation

BCR – B cell receptor

TCR – T cell receptor

APC – Antigen-presenting cell

Ig – Immunoglobulin

GBS – Group B *Streptococcus*

BAM – β-barrel assembly machinery

emPAI – Exponentially modified protein abundance index

T2SS – Type two secretion system

TMH – Transmembrane helix

Tat – Twin arginine translocation

PSSM – Position-specific scoring matrix

II Summary (German)

Impfungen haben die Welt verändert. In den letzten 200 Jahren konnten mit ihrer Hilfe eine Vielzahl von Krankheiten entweder unter Kontrolle gebracht oder ganz ausgerottet werden, darunter die Pocken, Masern, Diphterie, Polio und andere. Dennoch töten Infektionskrankheiten auch heute noch Millionen von Menschen jedes Jahr, und dies könnte mit Impfungen zumindest teilweise verhindert werden. Die Entwicklung neuer Impfstoffe mit klassischen Methoden ist Kosten- und zeitaufwendig, und nicht alle so entwickelten Impfungen erfüllen am Ende die Sicherheitsbestimmungen. Dank der Verfügbarkeit von kompletten Genomen vieler Krankheitserreger kann man heute auch alternative Wege beschreiten um Impfstoffkandidaten zu idnetifizieren, die sogenannte „reverse Impfstoffentwicklung“. Parallel dazu wurden immunoinformatische Methoden zur Identifizierung von Epitopen entwickelt, die für die Entwicklung von peptid-basierten Impfstoffkandidaten wichtig sind („epitope mapping“). In der vorliegenden Arbeit habe ich clustering-basierte Techniken der „reversen Impfstoffentwicklung“ mit Epitop-Identifizierungstechniken kombiniert, um konservierte und immunogene Bereiche in oberflächenexponierten Proteinen von Gram-negativen Bakterien zu finden, die für die Entwicklung neuer Impfstoffe geeignet sind.

Im Rahmen dieser Arbeit habe ich eine neue, präzise, Homologie-basierte Methode zur Vorhersage der Oberflächenlokalisation von Proteinen in Gram-negativen Bakterien und Archeen entwickelt, die sowohl auf neu sequenzierte Genome als auch auf massenspektrometrische Daten angewendet werden kann. Darüber hinaus habe ich eine Vorgehensweise zur Vorhersage immunogener Peptid-Epitope etabliert, die es erlaubt, B- und T-Zell-Epitope vorherzusagen, die innerhalb definierter Gruppen Gram-negativer Krankheitserreger konserviert sind. In einem weiteren Teil des Projektes habe ich den Einfluß bestimmter Aminosäuren sowie deren Position im C-terminalen Insertionssignal von bakteriellen Außenmembranproteinen untersucht. Diese Analyse zeigt die Existenz von Sequenzmotiven, die sowohl für taxonomische Gruppen als auch für Untergruppen von Außenmembranproteinen spezifisch sind. Des Weiteren haben diese Arbeiten Implikationen für die heterologe Expression von solchen Proteinen in *E. coli*.

II Summary (English)

Vaccination is a great success story. In the last 200 years it has controlled and eradicated a number of deadly diseases like small pox, measles, diphtheria, polio and others. But still, there are many infectious diseases which kill millions of people each year that could be controlled or extirpate with a vaccine. The discovery of vaccines by classical methods is costly, time-consuming and the results are not always completely safe to use. In the post-genomic era, the availability of complete genomes of pathogenic organisms has helped in indentifying surface-exposed proteins, which are potential vaccine candidates (‘reverse vaccinology’). In parallel, immunoinformatics techniques and tools have been developed to indentify immunogenic peptide epitopes from proteins of pathogenic organisms (‘epitope mapping’), which can be used to develop peptide-based vaccines. Here I have used a clustering-based reverse vaccinology method and combined it with epitope mapping techniques to indentify peptide epitope sequences that are conserved in surface-exposed proteins among Gram-negative bacterial pathogens, and that could be used in the development of new vaccines.

In this work, I established a highly precise consensus subcellular localization prediction pipeline for gram negative bacteria and archaea, including prominent pathogens, based on a clustering approach. This can be used to indentify surface exposed proteins of the pathogens, and to annotate subcellular localization of newly sequenced genomes of Gram-negative bacteria and archaea and of proteins identified in mass spectrometry experiments. I have also established an ‘epitope mapping’ pipeline, which can be used to identify the B cell and helper T cell epitopes conserved in different pathogenic strains of a species. As part of this work, I analyzed the influence of amino acids and their position in the C-terminal insertion signal of bacterial outer membrane proteins, revealing the presence of patterns, which are specific for both taxonomy classes and protein classes. Additionally, these results have implications for the heterologous expression of such proteins in *E. coli*.

III. List of publications and contribution to the publications

THEIN, M., SAUER, G., PARAMASIVAM, N., GRIN, I. & LINKE, D. 2010. **Efficient Subfractionation of Gram-Negative Bacteria for Proteomics Studies.** *Journal of Proteome Research,* 9, 6135-6147.

The initial idea was conceived by Dr. Dirk Linke. The workflow of the project, different fractionation methods and mass spectrometric data analysis were carried out by Dr. Marcus Thein and the mass spectrometric experiments were done by Dr. Guido Sauer. Iwan Grin helped in identifying proteins detected in the mass spectrometry data. I predicted the theoretical SCL for all proteins from the E. coli BL21 genome and also predicted the SCL of proteins identified from the mass spectrometry analysis.

PARAMASIVAM, N. & LINKE, D. 2011**. ClubSub-P: Cluster-based subcellular localization prediction for Gram-negative bacteria and Archaea.** *Frontiers in Microbiology,* 2:218.

The initial idea was conceived by Dr. Linke. I designed the workflow and carried out the complete project. I selected suitable tools for the experiment and wrote PERL scripts for the automatic submission of sequences to online servers when a standalone version of a tool was not available. The output was parsed and stored in MySQL databases and was later analyzed using PERL scripts. The precomputed results were then stored in MySQL and I integrated the database into MPI bioinformatics toolkit via Ruby on Rails web framework. I wrote the manuscript and Dr. Linke helped in improving it.

PARAMASIVAM, N., HABECK, M. & LINKE, D. 2012. **Is the C-terminal insertional signal in Gram-negative bacterial outer membrane proteins species-specific or not?** *BMC Genomics (Under Revision).*

The initial idea was conceived by Dr. Linke. I designed the workflow of the project and used different secondary structure prediction programs to obtain the C-terminal β strands. Then I used different statistical methods to cluster the obtained motifs. Michael Habeck suggested using the Hellinger distance and he also provided an initial script for the analysis. I updated the script and analyzed the data and found the presence of overlapping signals. I wrote the manuscript and Dr. Michael Habeck and Dr. Linke read and improved the manuscript.

IV. Introduction

IV.a. Bacterial Pathogenesis

We are frequently infected by innumerable pathogens, but there are various immune mechanisms to defend our body from these infections. Pathogenesis arises from the ability of these pathogens to disturb the immune system and thus, to cause diseases. Pathogens can be of four different types: bacteria, viruses, pathogenic fungi, and relatively large and complex eukaryotic organisms collectively called parasites [1]. Bacteria can cause many deadly diseases, and in the light of increasing antibiotic resistance world-wide, this is a re-emerging problem. Based on their degree of pathogenicity, bacterial species can be classified into three categories. Bacterial species that are primary cause of a disease in an individual with a healthy immune system are called primary or frank pathogens (*e.g*. *Salmonella sp.,*). Opportunistic pathogens (*e.g*. *Escherichia coli*) are species that are usually non-pathogenic, but can cause a disease in a favorable situation. Species that never or rarely cause disease are termed non-pathogens (*e.g*., *Lactobacillus acidophilius*) [2].

Pathogenic bacteria enter the human host mainly through the skin, the pulmonary tract, the digestive tract or the urinary tract. The squamous epithelial cells of skin and mucous epithelial cells of these tracts restricts the entry of pathogens into the host, and thus provide the initial defense mechanism against these pathogens. These membranes are coated with protective layers of lysozyme, *e.g*. lactoferrin and lactoperoxidase, which can kill microbes or restrict their growth [2]. These epithelial cells are also constantly replaced and shed along with the microbes attached to it. But many bacteria have developed specialized adhesins such as pili or fimbriae to attach strongly to the mucous membranes to start colonization [2].

Bacterial species secrete a variety of factors, which help them to adhere, colonize, and invade into host tissues, to obtain nutrients, and to protect them from host immune cells [2-4]. These factors, secreted or expressed by the bacteria for their survival, evoke a disease in the host; for this reason they are termed virulence factors. But, toxins are a subclass of virulence factors, which actively damage the host cells and/or tissues, these are toxins can be of two types, endotoxins (toxic substances released after the lysis of the bacteria, like lipid A moiety of outer

membrane LPS) and exotoxins (toxic substances that are secreted by the bacteria). The endotoxins are mostly structural elements in Gram-negative bacteria and induce a wide range of immune responses in the host, while the effects of exotoxins are very local and restricted to the cell type and their receptors [2].

IV.b. The host immune system

Generally speaking, the human immune system can be classified into an innate and an adaptive immune system. The innate immune system is non-specific and provides the initial defense against pathogens; it induces the more specific adaptive immune system *via* chemical signals and direct interactions. The main cellular components of the innate immune system are granulocytes (polymorphonuclear leukocytes (PMNs)) and macrophages [1]. If the pathogens manage to cross the barrier of the epithelial or mucosal membrane and start to infect tissues, they will encounter the first line of defense represented by innate immune cells such as macrophages, dendritic cells and neutrophils that reside there. These cells possess membrane receptors called pattern recognition receptors (PRR), which include the LPS receptor (CD14), Toll-like receptors, the mannose receptor and the scavenger receptor. These receptors are not very specific, but recognize common repeating patterns present in bacterial surface molecules, and engulf them into to a vesicle called the endosome which subsequently fuses with a lysosome to form the phagolysosome. The enzymes in the phagolysosome digest and kill the pathogens; this whole process is called phagocytosis, so these immune cells are also called phagocytes. During phagocytosis, these cells secrete cytokines which attract other immune cells to the site of infection and lead to an increased immune response called inflammation. Some bacterial species have developed resistance against the digestion by lysozyme by encapsulating themselves, which induces the immune cells to increase the inflammatory response to kill the bacteria. This increased inflammation can instead of destroying the microbes damage the host cells, leading to host-mediated pathogenesis [2-3]. The antigen-presenting cells (APCs, *e.g*. macrophages) of the innate immune system phagocytose the bacteria in the tissues, digest the pathogens and move to the lymph nodes, where they activate the more specific immune reaction against the pathogens.

The innate immune mechanism is non-specific, but is important for activating more specific adaptive immune mechanisms against the pathogen. The lymphocytes are the main players in the adaptive immune mechanisms, which are responsible for humoral and cytotoxic immune responses. In these processes, soluble antibodies are secreted by B cell lymphocytes that are responsible for the humoral immune response, while cytotoxic T cells or CD8 T cells are responsible for the cytotoxic immune response. Both of these types of lymphocytes depend on helper T cells or CD4 T cells for activation [5]. Before encountering antigens, lymphocytes are naïve and circulate between the blood and lymph vessels, but once they encounter pathogens, they proliferate and mature into effector cells. The lymphocytes have surface receptors that are very specific in their binding to pathogenic antigens or epitopes. The B cells directly recognize the epitopes present on the surface of extracellular pathogens and other soluble foreign bodies with their membrane bound receptors (BCR). In contrast, T cells can only recognize peptide epitopes displayed on the cell surface as by the major histocompatibility complex (MHC) molecules after digestion as part of the innate immune response. The BCRs are membrane-bound immunoglobulins (class IgM) which have a very specific antigen-binding site. The binding of BCR to antigens triggers the first set of activation signals to B cells; for complete activation, a second set of activation from activated helper T cells is necessary. The digested peptides in the phagolysosomes of an APC are attached to the MHC class II molecules and are displayed on the surface of the APC. In lymph nodes, the naïve helper T cells recognize the peptide/MHC class II complex on professional APCs, and in the presence of co-stimulatory molecules, become activated. When an activated helper T cell recognizes the same peptide/MHC class II complex on a B cell surface, it confirms that the epitopes recognized by B cells are foreign and activates the B cells by the second set of activation signals (cytokines) [5]. The activated B cell then proliferates and matures into effector plasma and memory B cells. The plasma cells secrete the soluble immunoglobulins (IgG) with the same antigenic specificity as that of the BCR. These secreted antibodies either bind to the exotoxins and neutralize them or bind to the endotoxins and opsonize the pathogen for phagocytosis. The memory B cells will live for longer time periods and get activated quickly without the secondary activation from helper T cells when they encounter the same antigen again in the future. Not all of B cell activations are T

cell-dependent: when the B cell receptors are provided with enough antigens like LPS or other toxins, B cells are activated directly.

There are two types of MHC molecules: MHC class I molecules, which are present in all nucleated host cells and display proteasome-digested peptides from intracellular pathogens and tumor antigens, and MHC class II molecules which are present on the surface of APCs and B cells and display protease-digested peptides from extracellular pathogens engulfed in vesicles. The MHC class I molecule is made of two polypeptide chains, the α-chain and β-2 microgobulin, where the α-chain consist of three domains. The α-1 and α-2 domains form the peptide-binding groove in MHC class I molecules, which is closed on both sides and can accommodate peptides of a length of 8-11 amino acids. The α-3 domain contains the transmembrane helix which anchors the molecules to the lipid membrane. The MHC class II molecules are heterodimers consisting of α and β chains, whose α-1 and β-1 domains form the peptide-binding groove which is open, so that peptides of a length of 10-18 amino acids can bind. The α-2 and β-2 domains consist of transmembrane domains which anchor the molecule to the membrane [1, 6]. Even though the peptide groove is open and MHC class II molecules bind to longer peptides, the actual peptide binding 'core' is only nine amino acids (AAs) long, but the flanking residues also influence the peptide binding.

The peptide/MHC complex is recognized by the T cell receptor (TCR) on the surface of T cells. The cytotoxic T cells recognize peptide/MHC class I complexes and helper T cells recognize peptide/MHC class II complexes. The activated cytotoxic T cells secrete perforins and other proteases, which form pores in the membrane of the targeted cells and thus trigger the apotoposis process of the cell. On the other hand, activated helper T cells won't kill any pathogens, but have an effector function of activating different immune cells for destroying the pathogen. The activated helper T cells are of two types: T_H1 which activate macrophages to kill intravesicular pathogens and T_H2 which recognize the peptide/MHC class II complex on B cells and activate the B cells.

The more specific adaptive immune system takes some days to weeks from the initial infection by microbes to produce antibodies, but the maturation of activated lymphocytes also produces memory B and T cells, which live for a long time period. When these memory cells encounter the same pathogen in the future, the secondary specific response against these pathogens will be much quicker compared to the

weeks in the primary response. Once the adaptive immune response is triggered against a pathogen, the resulted memory cells will provide lifelong immunity against the pathogen. This effect is the basis of all vaccines.

IV.c. Antibacterial drugs and vaccines

The bacterial infectivity results from the disruption of balance between the bacterial virulence and host immune resistance [2]. When a highly virulent pathogen disrupts the balance, various therapeutic antimicrobial drugs administered to regain the balance and protect the host from these pathogens. These antimicrobial drugs protect the host either by killing the microbes or by reducing the virulence by restricting their growth. Generally these drugs block the function of a targeted protein (typically an enzyme), which may be an essential protein, or it may be a protein in the pathway of pathogenesis, whose absence or blockage will decrease the virulence of the bacterium. Within a few years of introduction of all clinically used antibiotic, resistant bacterial strains have been reported and the widespread use of broad-spectrum antibiotics has led to the emergence of multidrug-resistant bacteria [7], which are a major concern. Antibiotic resistance is acquired by the pathogen in different ways: by acquiring mutations and altering the drug targets, by preventing the antibiotics from reaching the drug targets, or by producing enzymes (*e.g*., β-lactamase) that destroy or deactivate the antibiotics [3, 7]. Pathogens can also acquire antibiotic resistance genes present in the pathogenicity islands of other bacteria through horizontal gene transfer [3]. The recent emergence of multi-drug resistant Gram-negative pathogenic bacterial strains like *E. coli*, *Pseudomonas aeruginosa, Acinetobacter sp., Klebsiella pneumoniae, or Mycobacterium tuberculosis* poses a real threat to the treatment of infectious diseases globally [3]. We have to develop new therapeutic measures and strict clinical practices to control these multi-drug resistant strains.

Vaccination is a great success story. In the last century it helped to eradicate and control many deadly diseases including small pox, measles, polio and diphtheria [8]. Typical vaccines are dead pathogens, attenuated live pathogens, or parts of the pathogen which cannot cause the disease but induce an immune response that leads to adaptive immunity and production of memory B and T cells [9]. These memory cells protect the host from the subsequent infection by the pathogen. In the

conventional or first generation vaccines, pathogens were cultivated and various biochemical, immunological and microbiological methods were employed to find a suitable antigen from the pathogens [10]. There are many drawbacks in producing vaccines in the conventional way. Not all pathogens can be cultivated under laboratory conditions. Only highly produced proteins can be used in these studies, but most of the highly abundant proteins are not immunogenic. A high level of safety is required while cultivating the pathogens. Insufficient killing/attenuation in the vaccines may leak the pathogen into the healthy population. And finally finding a vaccine candidate is a very time-consuming procedure [11].

These disadvantages were reduced in the second generation of vaccines. With better understanding of the pathogenesis of different organisms, virulence genes were mutated or knocked out to generate attenuated vaccines. Recombinant DNA technologies were used to clone genes of pathogenicity factors and the products are used for subunit vaccine developments. Since these methods do not involve whole pathogenic organisms, these vaccines are safer than the previous generation of vaccines [11].

The availability of complete genome sequences of pathogens, combined with recombinant DNA technology, gave rise to the third vaccine generation, or reverse vaccinology methods [12]. In these methods, computational tools are used to identify surface-localized or secreted proteins from a proteome. These computationally predicted candidates are expressed and tested for immunogenicity and antibody production. The successful identification of a protective vaccine candidate against serogroup B *N. meningitidis* [13] using reverse vaccinology showed the power of genomic approaches in vaccine candidate identification [14]. This led to the broad use of this technology in identifying vaccine candidates for other human pathogens as well [14]. Applying reverse vaccinology to find a universal vaccine candidate against eight different strains of group B *Streptococcus* (GBS) [15], led to the finding that the genetic variability between strains of the same species is much larger than expected [16]. Only 80% of all genes found in the species are present in all the strains (termed as the “core” genome), and 20% of the genes were lacking in at least one of the strains (termed as the “variable” genome). The pan-genome (combining all the genes from all the strains) is much larger than the average genome of an individual strain, and mathematical extrapolation predicted that every time a new strain is sequenced, it adds 15 to 30 new genes to the pan-genome [8]. Such

species are said to have open pan-genomes. On the other hand, there are other species (*e.g*. *B. anthracis*) whose pan-genome can be described by a limited number of genomes (4 in the case of *B. anthracis*); these are termed as closed pan-genomes [16]. This demonstrates that genomes of multiple strains are necessary to understand the pan-genomic structure of a species. These led to the development of “pan-genome-based reverse vaccinology”, in which analysis and screening of the genomes of multiple strains of the species were used to discover conserved antigens, and this can be used to develop a universal vaccine for the species [15-17]. This approach can be further expanded to find conserved antigens in a genus, family or in an even broader taxonomic classification of pathogens. Peptide vaccines are based on recently developed immunoinformatics methods which search for B and T cell epitopes in whole pathogen genomes [18]. These tools can predict peptides that bind to MHC class I and II molecules (which are usually referred to as T cell epitopes) and continuous and discontinuous B cell epitopes. Following ‘epitope fishing’, various criteria are applied to select a set of optimal epitopes to develop antigen-specific epitope-based peptide vaccines.

V. Aims of the study

The main aim of this thesis work was to predict broad-spectrum vaccines against Gram-negative bacterial pathogens. I used a combination of clustering-based reverse-vaccinology and epitope-mapping methods to identify broad-spectrum vaccine candidates. To develop a broad-spectrum vaccine, a typical pan-genome based reverse-vaccinology method uses surface-localized proteins (outer membrane or extracellular proteins) conserved among different strains of a pathogenic species. However, we tried to apply a clustering-based reverse vaccinology approach to indentify clusters of surface-exposed sequences that are conserved in different pathogenic species. As an initial step, I set out to predict the surface-exposed proteins. Since individual predictors are not very precise, I wanted to combine different feature prediction and subcellular localization (SCL) prediction tools into a rule-based pipeline to get a highly precise consensus SCL prediction. Following this, I indented to cluster the sequences based on their sequence identity and select the surface-exposed proteins conserved in different pathogenic species for further analysis. In addition, I wanted to use the homology information derived from the clustering to improve the SCL predictions as well. We desire to develop a peptide-based vaccine, so, from the conserved surface-exposed outer membrane protein (OMP) clusters, I wanted to indentify the B cell and helper T cell epitopes using different immunoinformatics tools. There are various important factors to be considered while selecting the candidates for a broad-spectrum epitope-based peptide vaccine against extracellular Gram-negative bacterial pathogens, which I have considered using different filters while selecting the potential vaccine candidates.

Not all genes in a genome are expressed at all times. The expression of a gene depends upon the required cellular function in the given environment. Thus, proteins produced during the onset of infection are ideal candidates for vaccine development. We intended to use mass spectrometry to analyze the production of OMPs. We tried to establish an efficient and easy-to-use subcellular fractionation (SCF) technique which can be used to obtain and analyze outer membrane (OM) fraction from different Gram-negative bacterial pathogens. For this, we investigated five different cellular fractionation methods and selected the best method which maximizes the OMPs discovery with few non-OMP contaminations in OM fractions.

The β-barrel assembly machinery (BAM) in the OM of Gram-negative bacteria is responsible for folding and insertion of OMPs in to the OM. The BAM recognizes its substrate OMPs by their C-terminal β-strand. It has been reported [19] that this recognition is species-specific and there are also several reports that heterologous expression of OMPs in *E. coli* can be lethal. Since OMPs are important vaccine targets, their heterologous expression has much importance. Thus, I set out to find whether the proposed species specificity among the C-terminal insertion signals is true or not.

VI. Efficient subfractionation of Gram-negative bacteria for proteomics studies

VI.a. Introduction

The OMPs and extracellular proteins in the Gram-negative bacterial pathogens are the first to get in contact with host tissues and immune cells. The immune cells recognize these proteins and other toxins as foreign bodies and initiate an immune response to defend against infection by these pathogens. Since the OMPs include many virulence proteins, these proteins are potential vaccine candidates. But as not all the proteins in the genome of an organism are produced at the same time, gene expression is based on the required functionality at given environmental conditions [20]. So, identifying the complete set of proteins in pathogenic bacteria at an early stage of infection is very important for indentifying suitable vaccine candidates.

Mass spectrometry is a widely used proteomic technique to characterise proteins from different cellular compartments. It has been used to indentify novel vaccine candidates from various human pathogens [9]. We wanted to use the technique to characterise the set of immunologically important OMPs from Gram-negative bacteria. To perform a mass spectroscopy experiment, it is important and most crucial to purify and enrich the OMPs in OM fractions by avoiding or at least minimizing non-OMP contaminants [21].

To establish an efficient and easy-to-use protocol for subcellular fractionation (SCF) separating the OM components from rest of the cellular components for proteomics analysis, my colleagues (Dr. Marcus Thein and Dr.Guido Sauer) compared different SCF techniques (methods 1 to 5) [21]. Among these, methods 1, 2, and 3 focus on the fast enrichment of OMPs in the OM fractions, and methods 4 and 5 focus on the fractionation of proteins into cytoplasmic, IM, periplasmic and OM fractions. The detailed summary of these methods is given in the methods section of [21] under the subheading ‘Subfractionation methods’. Following fractionation by differential centrifugation, the proteins were solubilized and separated in one-dimensional SDS-PAGE, mainly to remove lipids and other membrane components in the process. Then the proteins were digested in-gel by trypsin and mass spectrometric analysis was performed using an ion-trap electronspray ionization-

mass spectrometry (ESI-MS) equipped with a nanoLC system for peptide separation. The data from the experiments were used in the Mascot software to identify proteins from the Swiss-Prot or non-redundant NCBI databases.

The best SCF method was chosen based on its ability to find more OMPs with less non-OMPs (contamination from other cellular compartments) in the OM preparation. For comparing different SCF techniques, we used *E. coli* BL21 to select the best SCF method and used it to obtain OM fractions from pathogenic *E. coli* 536, *E. coli* 2348/69, *Pseudomonas aeruginosa* PAO1 and *Yersinia pseudotubercolusis* IP32953.

VI.b. Prediction of outer membrane proteins from the *E. coli* BL21 genome

We used SCL prediction tools PSORTb [22], BOMP [23] and HHomp [24] to get a theoretical estimate of OMPs from *E. coli* BL21. The tool PSORTb, which includes different analytical modules, classifies or assigns the protein to one of the five different SCL in Gram-negative bacteria. It predicted 83 proteins, but PSORTb does not distinguish between OMPs and lipoproteins in its OMP prediction. HHomp identifies OMPs by detecting homology to known OMPs using HMM-based sequence similarity detection. It predicted 69 OMPs. BOMP combines two different components – the presence of a C-terminal β-strand and amino acid composition - to classify proteins into OMP or not. It predicted 73 OMPs. In total, these methods identified 121 different OMPs, but only 38 were found by all the methods. The numbers of overlapping and uniquely predicted OMPs were shown in Figure 1 of [21].

VI.c. Prediction of outer membrane proteins among proteins characterised from mass spectrometry

Five different SCF methods were compared to select efficient SCF techniques. Each of the SCF methods was performed three times to check the reproducibility of the methods. Combining the proteins identified from the 3 different trials, methods 1 - 5 found 256, 278, 413, 431 and 334 proteins, respectively in the OM fractions. Since PSORTb predicts SCL much faster than HHomp and BOMP, I used it to annotate the SCL of the proteins identified from mass spectroscopic experiments and it annotated 28, 36, 32, 37 and 37 OMPs from methods 1 – 5,

respectively. This is only 21% (method 3) to 41% (method 4) to the total proteins identified by the theoretical predictions. The rest of the proteins are non-OMPs from different subcellular compartments and the majority of them are cytoplasmic proteins. When PSORTb prediction scores are not above a threshold value, it does not assign localization to a protein. Instead PSORTb annotates it as 'unknown' and some of the proteins in the dataset were annotated as 'unknown' as well. In total, 44 unique OMPs were found from these five different methods to obtain OM fraction for mass spectrometric experiments. These unique proteins are only 53% of total OMPs predicted by PSORTb for the *E. coli* BL21 proteome. The list of identified OMPs from these experiments are listed in Table 2 of [21].

VI.d. Selecting the best subcellular fractionation method

The aim was to select the best method which can enrich most OMPs with least non-OMP contamination in the OM fractions. We used the emPAI scores (exponentially modified protein abundance index) to quantitatively evaluate the identified proteins. Proteins with higher emPAI scores have high abundance and are also highly reproducible in different trials of the each method. We also noticed OMPs have significantly higher emPAI scores than non-OMPs; Figure 4 in [21] shows the comparison of different thresholds of emPAI scores and the number of proteins identified from different methods with the corresponding threshold. We selected emPAI 0.25 as the threshold, because most OMPs were indentified at this threshold but included the least non-OMP. At this threshold, method 2 identified 73% (65/89) of the total OMPs found without a threshold, and found only 46% (192/418) of non-OMPs. This shows that the most of the non-OMPs are present only at a low abundance in the OM fractions. At this threshold, methods 2, 3 and 5 contained a higher percentage of OMPs to non-OMPs compared to method 1 and 4.

VI.e. Discrepancies between bioinformatics and proteomics

The five different methods 1 – 5 found 28, 36, 32, 37 and 37 OMPs, respectively. Methods 4 and 5 found the maximum number of OMPs but this is only 45% of the total OMPs predicted from the *E. coli* BL21 genome by PSORTb. This shows that at the given growth conditions, only few OMPs were produced by the bacteria, but it is possible that proteins with lower abundance were not detected by

mass spectrometry. In addition, one has to notice that there is a set of proteins annotated as ‘unknown’ by PSORTb in both whole genome predictions and proteins identified by mass spectrometry, so it is likely that this set of ‘unknown’ proteins may contain some true OMPs as well.

VI.f. Outer membrane proteomics of pathogens

These comparisons show that method 2 is more specific for isolating OMPs with less non-OMPs, so we selected this method for further experiments. Moreover, method 2 represents a fast and easy protocol and in addition it is also independent of lipid composition of membranes and vesicle density, which vary a lot among different Gram-negative bacteria. We tested the applicability of this method on different pathogenic bacteria by applying the method to obtain OM fractions from four γ-proteobacterial pathogens. The obtained OM fractions were applied to SDS-PAGE and the gel was digested with trypsin. Following the digestion, the peptides were analyzed using a nano liquid chromatography-ESI ion trap mass spectrometer and the Mascot software was used to identify the proteins from the mass spectrometric data. I used PSORTb to indentify the OMPs from the protein identified from the OM separations of these pathogens. Similar to the *E. coli* BL21 runs, the OM fractions also contained non-OMPs. 32 OMPs from *E. coli* 536, 27 OMPs from *E. coli* 2348/69, 27 OMPs from *Pseudomonas aeruginosa* and 20 OMPs from *Yersinia pseudotuberculosis* were indentified. The list of OMPs identified are listed in the Table 3 of [21].

VI.g. Conclusion

We compared five different SCF methods to find an efficient and easy-to-use method and we identified method 2 as the most specific method towards the enrichment of more OMPs with less contaminations (non-OMPs). We tested the practicability of this method to identify OMPs from different bacterial species by applying it to four γ-proteobacterial pathogens. Since proteins are produced based on the functionality required by the bacterium at the given environmental conditions, at maximum we identified only 45% of the OMPs from *E. coli* BL21. Some of the proteins annotated by PSORTb as ‘unknown’ might be OMPs as well. Since we cannot afford to miss potential vaccine targets, we need a better and more reliable

method to annotate the SCLs of proteins identified from mass spectrometric methods.

VII. ClubSub-P: cluster-based subcellular localization prediction for Gram-negative bacteria and archaea.

VII.a. Introduction

In Gram-negative bacteria newly synthesized proteins in the cytoplasm need to be sorted to their native location/cellular compartment in order to perform their respective functions. These targeted proteins have distinct sequence as well as structural features like secondary structures, signal peptides and amino acid composition which are used by various protein sorting machineries to recognize their substrates. Such protein sorting machineries can be broadly divided into 3 groups, 1) machineries that translocates the protein across the IM or integrate the protein into the IM (*e.g.*, Sec, Tat, SRP, YidC, Lol and Holin machineries), 2) machineries involved in the translocation of proteins across the OM or integration into the OM (*e.g.*, Type II Secretion system (T2SS), T4SS, T5SS, T7SS and T8SS), and 3) machineries tat export proteins directly to extracellular space or into host cells from cytoplasm (*e.g.,* T1SS, T3SS and T6SS) [25]. The SRP (signal recognition pathway) or YidC pathway integrates proteins with transmembrane helices (TMHs) into the IM [26]. The Sec and Tat machineries first recognize the proteins with general secretory system (Sec) SPs or twin-arginine translocation (Tat) SPs and then cleave the SP before translocating them across the IM [25, 27]. Proteins with lipoprotein SPs are recognized and their SPs are also cleaved and translocated across the inner membrane and subsequently their N-termini are modified and lipid anchored to the IM or OM by the Lol sorting system [28]. The OMPs have general or lipoprotein SPs and are hence translocated across the IM by the Sec or Lol translocation system. Following this, the BAM in the OM recognizes a C-terminal insertion signal in the OMPs and integrates them into the OM [29]. The T3SS, which injects proteins into the host cytoplasm, recognizes its substrate by an N-terminal signal peptide (called T3SS SP) [30-31].

Assigning the SCL to proteins in this post-genomic era *via* experimental methods is a time-consuming and expensive process. To deal with this challenge, various bioinformatics tools have been developed based on existing experimental data to predict the sequence features and annotate their SCL. These can be classified into two sets, 1) tools that predict only the features from the protein

sequences, and 2) tools that combine these feature predictions and annotate the SCL of the proteins. There are numerous tools currently available with varying prediction precisions, but it has been reported that the consensus prediction by combing different prediction tools decrease the false positive prediction rate and increases the overall confidence of the prediction [32-35]. During the work, I noticed the misannotation of start codons in proteins affects the N-terminal signal peptide prediction, further leading to wrong SCL prediction. At the single sequence level, misannotation of start codons are difficult to observe, but can be seen as very obvious gaps in the N-terminal region while performing a multiple sequence alignment of homologous sequences. This varied start codon annotation among orthologous sequences emerges from the use of different gene prediction programs on closely related organisms [36].

In this work, I developed a pipeline, Cluster-based subcellular localization prediction (ClubSub-P), to annotate the SCL of Gram-negative bacterial proteins, in which I used rule-based integration of consensus predictions from different sequence feature prediction tools. In addition, I have used SCL predictions from homologous proteins to further increase the confidence of the SCL predictions. To check the applicability, I have further used this strategy to predict SCLs of archaeal proteins. Moreover, I have shown that the use of consensus and homology information has increased the precision of ClubSub-P over the state-of-art SCL prediction tools.

VII.b. Subcellular localization prediction at the protein level

We have used 18 different tools to predict consensus SCLs of proteins from 607 Gram-negative bacterial and archaeal organisms. The various tools used are listed in the Table 1 of [37] and the genomes used in the study are listed in the Data Sheet S1 in Supplementary Material of [37]. For Gram-negative bacterial sequences I predicted features like SPs, TMHs and β-barrel domains, and applied protein sorting rules to predict the final SCL of the proteins. I used 10 different tools to predict five different SPs for Gram-negative bacterial sequences. LipoP [38] was used to predict the presence of lipoprotein SPs. Since the IM retention signals in lipoproteins are not properly established, classification of the proteins into inner or outer membrane lipoproteins was not possible. Consensus results from TatP [27] and TatFind [39] tools were used to predict the Tat SP. Predictions from SignalP

[40], Predisi [41], RPSP [42] and Phobius [43] were used to predict the general SP and the type III SPs were predicted from the consensus from the recently introduced tools EffectiveT3 [30] and T3SS_prediction [31]. PilFind [44] was used to identify the type IV pilin-like SPs. When more than one SP was predicted, I combined the predictions in the above mentioned hierarchy of SPs to annotate the consensus SP of the sequence. Since SPs contains a small stretch of hydrophobic residues, in many occasions they are mispredicted as TMHs. Therefore, combining the consensus SP prediction with the TMH predictions should reduce these misannotations [43]. The results from HMMTOP [45], TMHMM [46] and Phobius [43] were combined to predict the consensus TMHs, and the consensus SP cleavage site predictions were used to remove the TMH when it overlaps with the SP predictions. HHomp [24] was used to predict the β-barrel proteins. Since HHomp runs are time-consuming, one random sequence was selected from a cluster for HHomp runs when one of the sequences in the cluster had a positive OM prediction from CELLO [47] or PSORTb [22]. When HHomp predicted the sequence with more than 90% probability to be an OMP, then all the sequences in the cluster were annotated to have a β-barrel domain.

The consensus SCLs at the protein level were predicted by combining the consensus feature prediction based on protein sorting rules mentioned in Table 2 of [37]. The SCL of proteins are annotated as ‘inner membrane’, when they have at least one consensus TMH. The proteins are annotated as ‘periplasmic’, when they have a cleavable general or Tat SP without a TMH or a β-barrel prediction. The proteins with lipoprotein SP were annotated as ‘inner/outer membrane lipoproteins’. The ‘OMPs’ were identified by the presence of a cleavable SP and β-barrel predictions. The extracellular proteins were identified with the presence of a Type III or IV SP prediction or extracellular prediction from PSORTb. The proteins without any TMH or β-barrel or SP or extracellular prediction were annotated as ‘cytoplasmic’. Lastly, the proteins contradicting these rules were annotated as ‘unknown’. In addition to the classical SCL annotation in Gram-negative bacteria, I predicted extracellular proteins with TMHs and SP, and OMPs with lipoprotein SP.

VII.c. Subcellular localization prediction at the cluster level

In addition to the consensus SCL prediction at the protein level, I added SCL information from homologous proteins to overcome the misannotation of start codons and to further increase the confidence of the SCL prediction. To get the homology information, I clustered the sequences from 607 Gram-negative bacterial genomes based on their sequence identity. To determine the clustering parameters at which sequence identity the SCL information from homologous proteins can be transferred, I clustered proteins with experimentally verified SCL using CD-HIT [48] with various parameters. I used 8,227 proteins from experimental PSORTb database, and clustering at 40% sequence identity and 80% sequence coverage produced 1,023 clusters, of which 94.2% had proteins with the same SCL, 4.6% clusters had proteins with overlapping SCL, and only 1.2% of clusters had proteins with contradictory SCLs. So I applied the same parameters to cluster 1,911,760 sequences from 607 Gram-negative bacterial proteomes. As a result, 1,620,033 sequences were clustered into 174,028 clusters with two or more sequences (291,727 singletons were not used in this study). To annotate the SCL of the cluster, the SCL of the proteins in the cluster were averaged. Any SCL having more than 70% was annotated as the SCL of the cluster. However, if none of the SCLs were above 70%, the SCL of the cluster was annotated as ‘uncertain’. As a result I was able to annotate 1,500,778 sequences with a SCL, that is 78.5% of the sequences used in the clustering. A brief summary of the number of clusters and sequences belong to difference SCLs are given in the Table 4 of [37].

VII.d. Subcellular localization prediction for archaea

To test the applicability of the pipeline on a different group of organism, I applied the methods to 65 archaeal genomes. The membrane architecture of archaeal organisms is more similar to Gram-positive bacteria than Gram-negative bacteria, where the OM is replaced by a thick cell wall.

As there are not many specialized feature prediction tools for archaeal proteins, I used tools developed for Gram-positive bacterial sequences. In addition, PRED-SIGNAL [49], which was created to predict archaeal general SP prediction, and FlaFind [50], which predicts archaeal peripilin SPs, were used. The list of the tools used for feature prediction in archaea were listed in the Table 1 of [37]. Besides

replacing the type III and type IV SP predictions with archaeal prepilin SP prediction, a similar Gram-negative bacterial consensus SP pipeline was used to predict lipoprotein SPs or Tat SPs or general SPs. The consensus TMHs were assigned in a similar manner to that of Gram-negative bacteria.

The rules that are given in the Table 3 of [37] were used to predict the consensus SCL at the protein level. Proteins with Tat/general/prepilin SP/extracellular prediction from PSORTb were annotated as ‘secreted/extracellular’ and proteins with lipoproteins SPs were annotated as ‘lipoproteins’, and ‘cell wall’ proteins were indentified from the PSORTb predictions. Proteins with one or more consensus TMHs were annotated as membrane proteins and proteins without any SP or TMHs or β-barrel were annotated as cytoplasmic proteins. I used the same clustering parameters used in the clustering of Gram-negative bacterial sequences, to cluster 151,553 sequences from 65 archaeal genomes, which resulted in 22,184 clusters that have two and more sequences. Of these sequences I annotated SCL for 104,896 sequences, which is 69.21% of the sequences from 65 archaeal genomes. Similar to Gram-negative bacterial cluster-based SCL annotation, the SCL that was assigned to more than 70% proteins in a cluster was annotated as the SCL of the cluster. A detailed summary of the number of archaeal clusters and proteins annotated to different SCLs are give in the Table 7 of [37].

VII.e. Clustering-based comparison of signal peptide and transmembrane helix prediction tools

Due to the fact that most sequences with experimentally verified SCL were used in the training of the tools, it was difficult to find a golden dataset to test and select the high precision tools for consensus predictions. Thus, tools were selected on the basis of their consistency in the predictions of features among homologous sequences. As the sequences in the clusters were more than 40% identical with each other, applying a highly precise feature prediction tool should then return consistent predictions for all the sequences. Moreover, tools showing inconsistency in their prediction among homologous sequences were considered to be low precision tools. If a tool gave a positive prediction for only less than 20% of the sequences in a cluster, it was presumed that the predictions were false positives. If a tool gave negative predictions for only 20% of the sequences in a cluster, it was

presumed that the 20% are false negatives. Figure 4 of [37] compares the consistency of feature prediction tools over the clustered Gram-negative bacterial sequence data set. The plots includes separate comparisons of Tat, T3SS, general SP and TMH prediction tools, along with the consensus prediction of each features. It is evident from the figure that Tat and T3SS prediction tools have more false-positive predictions and the consensus prediction taken from these tools greatly reduces the false positives, but not completely. It could be inferred from the consensus prediction that most of the general SP predictions are not only consistent across the sequences in the clusters but also with each other's predictions. Among the transmembrane prediction tools, HMMTOP shows slightly higher false positives than other tools, but consensus TMH prediction greatly reduces it. These comparisons show that combining consensus predictions greatly reduces false positives arising from individual predictors. Moreover, when the test data set are not available or very small, the cluster-based consistency test across the larger data set helps to test the precision of tools.

VII.f. Database availability

We developed a MySQL database from the SCL predictions from Gram-negative bacterial and archaea proteins. The database, ClubSub-P, is integrated into the classification section of MPI bioinformatics toolkit [51]. Users can easily explore through the precomputed SCL for more than 600 genomes and can search the database for GI accession number and sequence header information. In addition, users can also BLAST query sequences against the database from which sequences with more than 40% sequence identity over a length of 75% will be returned as a hit and their cluster's SCL will be annotated as query sequence's SCL.

VII.g. Performance measure

We compared the performance of ClubSub-P against state-of-art SCL prediction tools PSORTb and CELLO. For the evaluation I obtained the sequences with experimentally verified SCL from Uniprot. However, to avoid biased performance measure I removed sequences that are more than 40% identical with PSORTb training data set. The resulting 171 sequences I then used to compare the performance of PSORTb, CELLO and the ClubSub-P Gram-negative bacterial

module. Since there are very few archaeal sequences with experimentally annotated SCL, I included the sequences from the PSORTb training data set, following which I used CD-HIT to reduce the sequence identity to below 40% among the sequences. Thus I obtained 252 sequences that were used to compare the performance of PSORTb and ClubSub-P over archaeal sequences. From the predictions, I calculated precision, recall, accuracy and MCC (Mathew's correlation coefficient) as given and described in the methods section of [37], and the results are given the Table 5 (Gram-negative bacteria) and Table 8 (archaea) of [37]. In the Gram-negative bacterial module, the performance measure shows that the overall precision and MCC of ClubSub-P (83.85%, 0.67) is slightly higher than PSORTb (80%, 0.59) and CELLO (66.67%, 0.6) with comparable accuracy between the tools. In the archaeal module as well ClubSub-P shows a slightly higher precision than PSORTb, but with slightly lower recall, accuracy and MCC. However, the presence of closely related sequences from the training dataset might bias the performance measures towards PSORTb, which cannot be avoided at this moment. But these results clearly show that the consensus SCL prediction and use of homology information have clearly increased the precision of ClubSub-P against the state-of-art SCL prediction tools.

VII.h. Incorrect start codons resulting in misannotated signal peptide

Along with the precision of the prediction tools, the quality of the input sequence is also an important factor influencing the quality of the output results; for example, proper start codon annotation is important for accurate N-terminal SP prediction. In a sequence cluster, if the majority of the sequences have a SP, the remaining SP-less proteins could either result from false-negative prediction from the SP prediction tools or they could also have an erroneous N-terminus resulting from a misannotated start codon. I looked for the SP-less sequences among secretory clusters ('periplasmic', 'lipoprotein', 'OMP' and 'extracellular with SP') in Gram-negative bacteria, and I found 3,558 SP-less sequences spread across 547 organisms. Previous studies [52-53] have shown that the misannotation of start codon in bacteria emerge from biased usage of uncommon start codons (UUG, GUG) in gene prediction methods. I noticed a biased frequency of uncommon start codons in the nucleotide sequences of these proteins. Among the normal sequences

the frequency of start codons AUG, GUG and UUG were 80.7%, 12.6% and 6.5% respectively. But among the SP-less sequences, the frequency of start codons AUG, GUG and UUG were 62.7%, 21.6% and 12.45% respectively. This clearly showed that misannotation of SP was largely due to the start codon misannotation rather than false-negative predictions from SP prediction tools. To confirm this hypothesis, I re-annotated these sequences with results from latest gene prediction tools like GeneMark [54], Glimmer [55] and Prodigal [56] and predicted SPs using signalP-HMM. From this analysis, I found 2,290 (64.4%) sequences with alternative start have a general SP prediction. It was clear that most of false-negative predictions of SP in the SP-dependent secretory clusters stem from start codon misannotation. Even though these protein sequences have a wrong open reading frame start, the homology information used for SCL annotation at the cluster level in ClubSub-P leads to annotating the proteins to the correct SCL without re-annotation of start codons.

VII.i. Conclusion

We have shown that the use of consensus prediction from multiple predictors reduces the prediction of false positives. I have also shown the use of simple homology information can considerably increase the precision of the predictions over the start-of-the-art tools and also overcomes the misannotation of start codons. In addition to classical SCLs, I have made more specific annotations like ‘OMP with lipid anchor’, ‘extracellular proteins with SP’ and ‘extracellular proteins with membrane anchor’ in Gram-negative bacteria and ‘cell wall proteins with membrane anchor’ and ‘extracellular proteins with membrane anchor’ in archaea. From these annotations, I have created a database, ClubSub-P, which can be used for SCL annotation of newly sequenced genomes. Importantly, the OMP and extracellular clusters can be used potentially to search for highly conserved vaccine candidates. In future, new tools and genomes can be easily added to the database and adding more specific feature prediction tools for different secretion machineries will help in identifying surface-localized proteins more accurately.

VIII. Is the C-terminal insertion signal in Gram-negative bacterial outer membrane proteins species-specific or not?

VIII.a. Introduction

The OMPs of Gram-negative bacteria are β-barrels composed of anti-parallel even-numbered β-hairpins arranged around a central pore. The precursors of these proteins are synthesized in the cytoplasm with an N-terminal signal peptide which is recognized and cleaved by the Sec machinery; the processed protein is subsequently translocated across the inner membrane also by the Sec machinery. Once the unfolded OMPs reach the periplasmic space, soluble chaperones like Skp, DegP and SurA bind to them, protecting them from aggregating and misfolding [57-58]. These chaperones deliver the unfolded OMPs to the BAM complex in the OM, which folds and integrates them into the OM [29]. The OMPs are delivered to the BAM complex *via* two separate chaperone pathways: one involves the protein SurA and the other comprises the proteins Skp and DegP. While the former works under normal conditions, the latter is activated under stress conditions [29, 59]. The central component of the BAM complex is BamA, which is a multi-domain protein composed of a 16-stranded β-barrel and five POTRA (polypeptide transport associated) domains [57]. The BAM complex also comprises accessory lipoproteins, BamB, BamC, BamD, BamE and the recently characterized BamF [57, 60]. The BAM complex recognizes its substrates, unfolded OMPs, by their amphipathic C-terminal β-strand (C-terminal insertion signals), but the exact structural mechanism of the recognition is still under investigation [19].

The expression of neisserial OMPs in *E. coli* was shown to have lethal effects [19] and it was hypothesized that inadequate recognition of neisserial OMPs by the *E. coli* BAM complex might be the reason for this lethality. The authors tested this hypothesis in *in vitro* by electrophysiological experiments; they checked whether the neisserial PorA and the synthetic C-terminal insertion peptide open the *E. coli* BamA channel in an artificial lipid bilayer. And indeed the neisserial PorA and its C-terminal insertion peptide did not open the BamA channel in the artificial lipid bilayer. The neisserial PorA, like *E. coli* PhoE, has a Phe at the C-terminal end; however, while the neisserial PorA has a Lys at the +2 position, the *E. coli* PhoE has a Gln. Further comparison between *E. coli* and neisserial OMPs revealed a strong preference for

positively charged amino acids at +2 positions in *Neisseria*. When the authors replaced Gln with positively charged amino acids in the *E. coli* porin PhoE, it failed to open the BamA channel, and when they replaced Lys with Gln in neisserial PorA, it activated the BamA channel. This led to the conclusion that the C-terminal insertion signal in OMPs is species-specific and that the +2 residues are responsible for this specificity. But this conclusion was arrived at based on comparisons carried out between two organisms and very few sequences, so it remained unclear whether this hypothesis holds true for all the Gram-negative organisms. Since we are interested in finding vaccine candidates among OMPs, their heterologous expression in model organisms have important biotech applications. I therefore used computational tools to investigate the proposed species specificity of these C-terminal insertion signals.

VIII.b. Extraction of C-terminal β-strands from 607 Gram-negative bacteria

I extracted the C-terminal β-strands from OMPs predicted in ClubSub-P [37], and also from the OMPs predicted among the singletons (which were obtained during the clustering of the sequences in ClubSub-P). I used the β-barrel topology predictions from ProfTMB [61] and the secondary structures predictions from PSIPRED [62] in HHomp [24] results in extracting 25,454 C-terminal β-strands, which were 10 to 21 AAs long. I then used the gapped-motif discovery tool, GLAM2 [63], to obtain statistically significant motifs 10 AAs long from these β-strands; all motif instances with gaps were eliminated. In this data set, 437 organisms had more than 20 OMPs which constituted 22,447 C-terminal β-strands. I used these motifs instances for further studies. These motif instances can be classified into different classes or groups based on the taxonomic classification of the organisms, and also based on the number of β-strands present in the original OMP.

VIII.c. Sequence-based and PSSM-based clustering

To check whether these C-terminal insertion signals are species-specific or not, I clustered the motifs on an individual motif level and also at the organism level. If the motifs are species-specific, then these sequences should cluster according to taxonomic class and organisms should group together based on their taxonomy classifications.

The sequence-based clustering in CLANS using the PAM30 scoring matrix did not separate the motifs into different taxonomic classes, but there was a very crude separation based on OMP classes.

I then generated a PSSM matrix for each organism as described in the methods section of [64] and clustered them using hierarchical clustering. The resulting dendrogram showed a good separation of organisms based on their taxonomic classes. But, when I used the R package pvclust [65] to access the uncertainty in the hierarchical clustering, I found that the AU (Approximately Unbiased) *p*-value and BP (Bootstrap Probability) values, which are used to measure the certainty of the clustering, were not correlated. So, I could not confirm or deny the species specificity of the C-terminal insertion signal with this clustering data.

VIII.d. Chemical descriptor-based clustering

These hierarchical clustering results suggest that there is a weak organism-specific signal, but the sequence-based clustering results suggest a weak OMP class-based clustering. So I hypothesized that the C-terminal insertion signals from different organisms overlap. If there is a complete overlap of C-terminal insertion signals between two organisms, then the BAM complex of the host organism will recognize all the OMPs from the other organism and *vice versa*. And if there is no overlap then none of the OMPs will be recognized for heterologous expression in the host organism.

I tested this hypothesis by clustering the organisms based on the pairwise overlap of peptide sequence space calculated from C-terminal insertion signals. Initially, the peptide sequences were represented using a five-dimensional chemical descriptor, which are the first five principal components derived from 26 different chemical and physical properties of AAs [66]. This resulted in a 50-dimensional peptide vector. Since the dimensionality of the data should be less than the sample size (minimum 21 sequences per organism) for further statistical analysis, the peptide vector dimensions were reduced to 12 using principal component analysis (PCA). As described in the method section of [64], all the 12-dimensional peptide vector for C-terminal β-strands from individual organisms were combined into a matrix. Then I calculated the mean and covariance for the matrices to fit a multivariate Gaussian distribution; this we call ‘peptide sequence space’. Then I used

Hellinger distance, a statistical theory method, to calculate the pairwise overlap between the multivariate Gaussians distributions from different organisms. The pairwise measures were used in CLANS to cluster the organisms.

In the cluster map (Figure 1A, [64]), each node is an organism, and the darkness of the edges connecting the nodes is directly proportional to the overlap of peptide sequence spaces between the organisms. The organisms are colored based on their taxonomic classes and one can notice that the organisms were crudely clustered based on the taxonomic classes. In the cluster map, *E. coli* strains are clustered together among other γ-proteobacteria and *Neisseria* species are clustered along the periphery of β-proteobacterial cluster, and a few α-proteobacteria are clustered between the β- and γ-proteobacteria clusters. Interestingly, the *Helicobacter pylori* strains clustered separately from the rest of the organisms, suggesting that they have a very distinct motif from rest of the organism's C-terminal insertion signals.

This clustering analysis agrees with the presence of an organism-specific signal, but a closer look at the cluster map shows that organisms with fewer OMPs are seen in the periphery of the cluster map and those with a larger number of OMPs are in the middle of the cluster map. A control experiment was carried out to check whether the observed organism-specific signal is true or if it is an artifact arising from the number of OMPs present in an organism. The positions of AAs in the motif were randomly shuffled and the clustering was done as described earlier. In the resulting cluster maps, organisms clustered as concentric circles when I colored them based on their taxonomic classes (Figure 2A of [64]). I noticed that the taxonomic clustering observed in Figure 1A was lost. And while I colored the organisms based on the number of OMPs present (Figure 2B of [64]), I noticed that the circles are formed by organisms with a similar number of OMPs, and it increases from periphery to the center. This confirms the presence of an organism-specific signal among C-terminal insertion signals from different organisms, which is lost when the AA positions are shuffled. Thus, the AA positions in the C-terminal insertion motifs are an important constituent of the signal.

VIII.e. High preference of residues at different positions

Previously, [19] suggested that the high preference for positively charged residues at the +2 position in neisserial C-terminal signals are responsible for the inadequate recognition of neisserial proteins by the *E. coli* BAM complex, but they compared only few sequences from these organisms. Therefore, I used more sequences from these organisms to check the frequency of residues at the +2 position. And I also noticed the high preference of positively charged residues (Arg, Lys) at the +2 position in neisserial C-terminal insertion signals. To check whether there is any general pattern present among the proteobacteria, I compared the frequency of positively charged residues at the +2 position from 437 Gram-negative organisms. Figure 4 of [64] shows that more than 60% of the C-terminal insertion signals from *Neisseria* species have positively charged resides at the +2 position. The frequency of these organisms stands out from rest of the organisms and even from closely related β-proteobacteria. I also noticed that 25 to 40% of *E. coli* C-terminal insertion signals have positively charged residues at the +2 position. Hence, the *E. coli* BAM complex should recognize C-terminal insertion signals with positively charged residues at the +2 position and in fact there is also experimental evidence for *E. coli* BAM complex recognizing OMPs with positively charged residues at the +2 position [67]. This suggests that the observed species specificity is not depending on positively charged residues at the +2 position.

The *Neisseria* porin PorA used by [19] for the investigation has a His at the +3 position, which is unusual for this position. The +3 position is usually occupied by a hydrophobic residue. Therefore I checked for the frequency of His at the +3 position in the C-terminal peptides of all the organisms. Figure 5 in [64] shows the high preference of His at the +3 position among β-proteobacteria, and the frequency in *E. coli* is almost zero. Further investigation revealed that the high preference of His at the +3 position in β-proteobacteria is mostly limited to 16-stranded porins, and structural investigation shows the His residues are present in the trimerization interface. The complete absence of His at the +3 position in *E. coli* C-terminal insertion signals might be the reason for the *E. coli* BAM complex not recognizing the neisserial PorA with His at the +3 position.

In the cluster map, Figure 1A of [64], the *H. pylori* strains are clustered separately from rest of the organisms, which suggests that *H. pylori* strains have a

distinct C-terminal insertion signal. I did not notice a high preference for any residue at the +2 position in the C-terminal insertion signals of *H. pylori*, but there was a high preference for Tyr at the +3 position, which is common among C-terminal insertion signals. In addition, I noticed a high preference of Tyr at the +5 position. Hydrophobic AAs are common at this position, but aromatic hydrophobic AAs are not. Therefore, I checked the frequency of aromatic hydrophobic residues at the +5 position in the C-terminal insertion signal from all proteobacteria. From the resulting percentage plot (Figure 8A and 8B of Paramasivam *et al*. (2012)), I noticed that the high frequency of Tyr is only seen among *H. pylori* insertion signals and it is uncommon among other proteobacteria, which includes *E. coli*. It has been reported earlier [68] that the expression of *H. pylori* OMPs in *E. coli* is lethal to the host. The expression of *H. pylori* OMPs were normal until they were translocated across the IM to the periplasm, where they started to aggregate. But the expression of *H. pylori* OMPs without a C-terminal is tolerated and there was no aggregation in the periplasm. The authors [68] concluded that the mis-targeting of *H. pylori* OMPs is lethal in *E. coli.* I hypothesize that the C-terminal insertion signals from *H. pylori* OMPs are not recognized by the *E. coli* BAM complex; therefore the OMPs are not inserted into the OM and start aggregating in the periplasm. Since the Tyr at +5 in *H. pylori* C-terminal insertion signals is uncommon in other proteobacteria, this might be the major reason for the inadequate recognition of *H. pylori* OMPs by the *E.coli* BAM complex.

VIII.f. Outer membrane protein class-specific and taxonomy class-specific signals

During these analyses I noticed a high prevalence of 16-stranded OMPs in some β-proteobacteria organisms and 22-stranded OMPs in some α-proteobacteria organisms. Since, 33.8% of OMPs are not classified under any OMP class, it not feasible to reduce the over-representation based on the OMP class alone. I therefore did a control experiment to demonstrate the influence of the over-representation of OMP classes in the clustering of organisms. I removed C-terminal insertion signals from one of each of the OMP classes (8-, 12-, 16- and 22-stranded) from organisms and clustered the data set. In the resulting cluster maps, Figure 9A and 9B of [64], the 8- and 12-stranded OMP classes were removed respectively. Since these classes do not have a high prevalence in any organisms, their absence did not affect the clustering. However, the OMP classes 16-stranded (Figure 9C, [64]) and 22-

stranded (Figure 9D, [64]) are over-represented in β- and α-proteobacteria, respectively, and their removal affected the clustering and the organisms were not clustered based on their taxonomy. This demonstrates the influence of the high-prevalence of OMPs in the clustering. I did not remove the over-represented OMPs from the data set, because all the over-represented C-terminal insertion signals still represent the true peptide sequence space of the organisms.

I also examined whether there is any signal arising from OMP classes that influences the clustering. To perform this test, I created two data sets, one containing organisms from all the taxonomy classes but with C-terminal insertion signal only from 22-stranded OMP class, and the other data set containing multiple representatives of organisms from γ-proteobacteria class, each representative containing C-terminal insertion signals from one particular OMP class. The clustering was done as described before and the cluster map obtained from these data sets are shown in Figure 10 of [64]. In the Figure 10A of [64], data set created from 22-stranded OMP class were clustered. In this cluster map the peptide space of the organisms were calculated only from C-terminal insertion signals of 22-stranded OMPs, and they clustered based on taxonomic classes. And in the Figure 10B of [64], data set from γ-proteobacteia alone was used to cluster. And organisms with more than one OMP class have multiple representative in the cluster map. But they clustered based on different OMP classes not on taxonomic class as the first data set. This demonstrates the presence of an OMP class-based signal along with the organism-specific signal in the C-terminal insertion signals.

VIII.g. Conclusion

I have showed that the prevalence of positively charged amino acids at the +2 position in neisserial OMPs is higher compared to other Gram-negative bacteria. Since there is experimental evidence about the functional expression of OMPs with positively charged residues at the +2 position in *E. coli*, the observed species specificity of the C-terminal insertion signal may not be due to the presence of a positively charged residue at the +2 position. However, the presence of His at the +3 position in the neisserial porin PorA and its complete absence in *E. coli* C-terminal insertion signals may be the reason for the inability of the *E. coli* BAM complex to recognize *Neisseria* OMPs. The chemical descriptor-based CLANS clustering

demonstrated the presence of organism-specific signals and I have also demonstrated that the proportion of OMP class present in an organism influences the average motif from an organism. These results confirm the presence of two overlapping signals, one based on OMP class and another based taxonomy class. Since most of the C-terminal insertion signals are very similar, the heterologous expression of proteins should be possible in most cases. While performing heterologous expression, it is better to check for the presence of individual insertion signals in the host proteins, rather than comparing the frequency plots from the two organisms. Since the clustering is based on entire sequence spaces, I could not confirm which residues or positions are important for the observed signals. However, once the C-terminal insertion signal binding site is known, one could study the co-evolution of the interacting residues. This will demonstrate the relative importance of residues and positions in the C-terminal insertion signal.

IX. Broad-spectrum epitope-based peptide vaccine candidates

IX.a. Introduction

The aim of vaccination is to provide lifelong immunity against pathogenic organisms; thus vaccines are aimed to induce specific adaptive immunity which comprises both humoral and cell-mediated immune responses. Since both B cell and helper T cell-based immune responses are needed to protect against an extracellular bacterial infection, a protective peptide vaccine should contain B cell epitopes and helper T cell epitopes (peptides binding to MHC class II molecules). Currently, our aim is to develop a broad-spectrum vaccine against extracellular Gram-negative bacterial pathogens, so we need to identify the peptide epitopes that are conserved in different strains of a species, genus or a family.

Typically, B cells epitopes are the discrete surface exposed regions of antigenic proteins to which the variable region of an antibody binds [18]. Most of these B cell epitopes are discontinuous at the sequence level, but are brought together by protein folding. B cell epitopes can, however, also be continuous, where an antibody binds to a surface-exposed, linear stretch of sequence of the antigen. The most important factor to consider when designing a peptide-based vaccine is the 'cross reactive immunogenicity' of the peptides, that is, the ability of the anti-peptide antibody to specifically bind to the antigen protein from which the peptide or epitope is derived [69-70]. The difficulty in characterizing non-linear epitopes of the complementary antigen surface and poor understanding of the recognition properties of cross-reactive antibodies are among the important reasons why B cell epitope prediction has lagged behind the T cell predictors in accuracy and reliability [18, 71]. The B cell epitopes should be surface-exposed for antibodies to bind, and the antigenic proteins should have a high expression level and should ideally be concentrated on certain patches of the cell surface for opsonization [72]. The precision of currently available discontinuous B cell epitope prediction tools is low and delivery of discontinuous epitopes in a native-like confirmation in vaccines is a challenging task [73]. Thus, in this study I have predicted only continuous B cell epitopes that are surface-exposed. I have used the programs BCPREDS [74] and BepiPred [75] to predict continuous B cell epitopes.

T cell epitopes are linear peptides displayed on MHC molecules that are derived from pathogens or tumor cells. The MHC loci are highly polymorphic, resulting in thousands of alleles with different peptide binding abilities, and the set of alleles present differs between individuals and also between different populations [76]. This makes individuals and populations react differently to the same pathogen. Thus, vaccines which include T cell epitopes should consider the MHC allele distribution in the target population. Peptides that bind to multiple alleles and cover a major percentage of the targeted population are better vaccine candidates. There are various allele-specific prediction tools available, but most of these predictors cover a limited set of HLA alleles for which more quantitative peptide binding data are available [77]. Since MHC molecules can be clustered into supertypes based on their binding specificities [78], various pan-specific HLA binding methods have been developed which indentify peptide binding to hundreds of alleles. These methods use experimental peptide binding data from multiple 'source alleles' to predict peptide binding to 'target alleles' in the same supertype for which no peptide binding data is available [77]. Compared to allele-specific methods, pan-specific methods have more allele coverage and reasonable prediction accuracy. MHC class I molecules have a closed peptide-binding groove with a defined binding site, but the peptide-binding groove of MHC class II molecules is open, so that the length of the peptides binding varies between 10 and 18 [6]. Even though the peptide binding 'core' is usually nine AAs long, the open ends of MHC class II molecules makes it difficult to predict the optimal binding core in peptides with high accuracy [77]. In this study, I have used the pan-specific method NetMHCIIpan-2.0 [79] to predict peptides binding for 639 HLA DR alleles and allele-specific method NetMHCII [80] to predict peptide binding for six HLA DQ and six HLA DP alleles.

IX.b. Epitope prediction pipeline

Using the ClubSub-P database that contains 607 Gram-negative bacterial proteomes, I identified 22,548 OMPs in 3,315 clusters. 1,290 of these clusters have more than five sequences, and I selected 119 clusters from these which have more than 70% of their sequences from pathogenic organisms and which are conserved across different pathogenic organisms from a broad range of taxonomic classification. These clusters were used to search for peptide epitopes that are

conserved among the different sequences in the cluster, so that these epitopes can induce a broad spectrum of immune responses against all organisms in the cluster. The prediction of epitopes was done using a random sequence from each of these clusters and conservation of the predicted epitopes was examined later at the cluster level.

I used BCPREDS [74] and BepiPred [75] to find continuous B cell epitopes and NetSurfP [81] to predict the surface exposure of the residues in the sequences. PSIPRED [62] was used to predict the secondary structure of the proteins and PROFtmb [61] was used to predict the β-barrel topology of the representative sequences. I used multiple sequence alignment (MSA) of the clusters to calculate the AA conservation in each column and the columns that had more than 80% identity were annotated as conserved.

All the predicted features were combined for each sequence and represented as a MSA and a sliding window of 10 AAs length was used across the MSA to find continuous B cell epitopes. If the sequence within the sliding window should be considered as B cell epitopes, it should pass some filtering criteria, which include, 1) 80% of the residues should have positive predictions from both continuous B cell epitope predictor programs, 2) 70% of the residues should have surface-exposed residue prediction by NetSurfP [81], 3) 70% of the residues should have loop structure predicted by PSIPRED [62] and 4) 70% of the residues should have outer surface coils annotation from PROFtmb [61]. These above criteria ensure that the B cell epitopes are present on the extracellular side of the membrane, so that the anti-peptide antibodies will have access to the epitopes on the surface of the proteins. If 80% of the AAs in the sliding window are conserved in the cluster along with above criteria, I annotated the residues in the sliding window as a potential B cell epitope. Then the sliding window was moved to find the next B cell epitope, and if overlapping windows were positively predicted, they were merged into one. Thus I obtained highly-conserved surface-exposed continuous B cell epitopes.

From the above analysis, I found 90 continuous B cell epitopes from 41 clusters, which are conserved among different taxonomy levels, in *Acinetobacter, Neisseria*, *Brucellaceae* and *Enterobacteriaceae*. To demonstrate our vaccine prediction pipeline, I selected the seven OMP clusters conserved among *Neisseria* for further analysis. These seven neisserial clusters contains 17 continuous B cell epitopes. I used the seven representative sequences from these clusters to predict

pan-specific MHC class II binding peptides using NetMHCIIpan-2.0 [79] and NetMHCII tools [80]. I selected helper T cell epitopes from the same protein to ensure that both B cell and T cell epitopes in the vaccine are also expressed in the pathogen at the same time. From the NetMHCIIpan-2.0 [79] and NetMHCII [80] results I selected 'strong binders', *i.e*., peptides with a high affinity (K_D < 50 nM) to MHC molecules. Since single mutations in the 'binding core' of the peptide alter the binding affinity greatly, only peptide 'binding cores' that are completely conserved in the cluster were selected. Since I do not currently have a target population, I ranked these completely-conserved peptides based on their ability to strongly bind to multiple HLA alleles, and the top 20 peptides that bind to more than 50 alleles were selected for each cluster. These peptides were then searched using BLAST (PAM30 matrix, e-value 100) against the human protein database for matches. The peptides that had matches were discarded from the data set because of possible cross-reactivity with human proteins. The top five MHC class II molecule binding peptides are reported in Table 1 along with the conserved continuous B cell epitopes.

IX.c. Conclusion

In this work I have applied the clustering-based reverse vaccinology to select OMP clusters from pathogenic Gram-negative bacterial organisms and used 'epitope mapping' techniques to select B cell and helper T cell epitopes from these OMP clusters. I have demonstrated our computational vaccine prediction strategy by predicting peptide epitope vaccine candidates from *Neisseria*. As a result, I have predicted 17 continuous surface exposed B cell epitopes from seven neisserial OMP clusters. Since we do not have any target population at this stage, peptides binding to maximum number of HLA alleles were selected as helper T cell epitopes from these clusters. As only the core of the helper T cell epitopes are reported here, flanking residues can be added to these peptides which may produce overlapping helper T cell epitopes. Further experiments have to be done to test both the production of these proteins in the pathogens, and the immunogenicity of the epitopes. The peptide epitopes identified from this analysis can be linked together using linker regions to develop multi-subunit vaccines.

Table 1: List of conserved surface-exposed B cell epitopes and MHC class II binding peptides from Neisseria OMP clusters

ClubSub-P cluster	Representative sequence (GI number)	Protein names	Gene Ontology (Molecular function)	B cell epitopes	Helper T cell epitopes	
					MHC class II binding peptide 'core'	# of HLA alleles it strongly binds
41572	194100067	Putative uncharacterized protein	None	VGGAVNNAA	YGRARALLA VSLALSHDK	337 91
47843	121634304	Putative hemolysin activator	None	AGGKTTGKY	LSRLQKAAQ ILIVRGYLT YQSSLAAER FYVSYGRGL FNNKFPLYR	246 217 174 160 135
46994	59802406	Putative adhesin penetration protein	serine-type endopeptidase activity	AGTNGHPYGGDY DEDEPNNRES GEKDATKTNG TRNATQNGN	IRFSPAYLA IRRRVLHYG YGIQARYRA FSVVSRNGV GLAFNRYRA	429 299 235 190 176
41006	121634057	Outer membrane protein OMP85	None	DPRKASTSIKQYK	MKFSYAYPL IQITPKVTK YYSATHNQT IIKSLYATG LRASRSKTT	322 218 217 189 163
47148	254804571	Lactoferrin binding protein A	Transporter activity	QPLSAEEEA GGGRILPDPMDYR YGTDEADKF NEAYSDNWA KPKSVSNRP	LKYSRTKFI YNRIKPKSV YLNLKKRLT VRAAKVGRR YRYVTWESL	386 328 271 135 128
41069	121634766	Putative outer-membrane receptor protein	Transporter activity	NYYNHPLPD	YRVFAQNKL FKPTPRYRI IRGQTGRRI WQKSLINKR IIWQKSLIN	328 324 292 252 229
47449	121634205	Putative ferric siderophore receptor protein	Siderophore transmembrane transporter activity	YDSQGYATAFGPKD QPASFAQTI TGSYDSRTQGMT TLRIPNPAAKARA	LNASAAVYR LRTVNAAFT YRARKNNLA YLATRFRAA VYRARKNNL	225 213 198 180 159

X. Conclusion and Outlook

In this work I have used 18 different secondary structure and SCL prediction tools to develop a pipeline for consensus SCL prediction of protein sequences from Gram-negative bacteria and archaea. On top of this I used homology information to overcome the misannotation of start codon and to further improve the confidence of the predictions. I was able to show that the use of consensus methods increases the consistency of the predictions in a cluster by cluster-based comparison of tools. In addition, I demonstrated that by using the consensus from different prediction tools and by use of homology information, the precision of the predictions is increased. I have applied this pipeline on a large scale and predicted the SCL for 607 Gram-negative bacterial proteomes and 65 archaeal proteomes. The precomputed database, ClubSub-P [37], created using the pipeline is available online in the MPI bioinformatics toolkit [51]. The database can be searched to annotate the SCL of proteins identified *e.g*. by mass spectrometry or genome sequencing. Additionally, the OMP clusters conserved in different pathogenic strains can be used to indentify vaccine candidates. The established clustering parameters can be used to cluster proteins from selected groups of pathogenic organisms and the SCL of the clusters can be annotated by querying a representative sequence against ClubSub-P. It is also possible to predict more specific SCL by adding more specialized feature prediction tools, *e.g.* when new tools are available for the different secretion machineries. Similarly, when new strains of Gram-negative bacteria and archaea are sequenced, they can be integrated in the clustering process, which may decrease the number of singletons and increase the percentage and quality of SCL annotations in all the sequences further.

I have used the OMP clusters from ClubSub-P from pathogenic organisms to identify new vaccine candidates. I created a pipeline to identify continuous B cell epitopes and MHC class II binding peptides (helper T cell epitopes) and demonstrated its functionality using OMP clusters conserved in *Neisseria*. We intend to indentify such highly conserved peptide epitope candidates from ClubSub-P OMP clusters conserved in different pathogenic organisms. In the context of ‘epitope mapping’, new tools will be employed to discover discontinuous and discontinuous B cell epitopes from clusters. In cases where vaccine discovery is targeted against a disease which is prevalent in a particular population, HLA allele distribution in that

population can be used to select relevant peptides that bind to HLA alleles prevalent in the population. Experimental microarray data sets of gene expression from pathogenic Gram-negative bacterial organisms available in public databases will be used along with the mass spectrometric data sets to identify vaccine candidates from the computational pipeline in the future.

Since we are interested to discovery peptide epitope vaccine candidates from the OMPs, the heterologous expression of these proteins in *E. coli* are important. However, it has been reported that the recognition of OMPs by the *E. coli* BAM is *via* a species-specific motif, but the comparison was performed with less than 25 peptides from two organisms. Therefore, I tested whether this holds true or not among all the Gram-negative bacterial species. We have used the OMPs predicted from ClubSub-P (including singletons) to obtain 22,447 C-terminal β-strands from 437 Gram-negative bacterial species. I have used chemical descriptors to define the peptide sequence space for each organism and used a statistical theory method (the Hellinger distance) to calculate the overlap of peptide sequence spaces between organisms and used this pairwise distance to cluster all organisms. From this analysis I have concluded the presence of two different signals relevant for proper OMP recognition by the BAM machinery: one is based on taxonomic class and the other on the OMP class. Thus, evolutionary distance between organisms plays a role for heterologous recognition, but also the proportion of proteins from different OMP classes present in an organism determines the average signal from an organism. I have also demonstrated that the proposed species specificity is not only determined by residues present in the +2 position, but instead varies among different organisms. I have also predicted that the high preference for Tyr at the +5 position in *H. pylori* C-terminal insertion signals is responsible for the inadequate recognition of these peptides by the *E. coli* BAM machinery. Furthermore, wet lab experiments have to be done, *e.g*. by mutating these residues and subsequent functional expression of *H. pylori* OMPs in *E. coli*, to prove the importance of these residues.

In a much broader sense, in this work I have created ClubSub-P, a database of precomputed SCL predictions which will be used to identify potential vaccine candidates from the outer membrane and extracellular protein clusters from the pathogenic organisms. Since the precision of SCL predictions of ClubSub-P are better than the state-of-art tools, it can be used in the functional annotation pipeline of newly sequenced bacterial and archaeal genomes. The concept of comparison of

performance of different tools based on the consistency across the sequences in a cluster, which was first used in ClubSub-P, can be adapted to many feature prediction programs. In this work we have used a clustering-based reverse-vaccinology approach to find potential broad-spectrum peptide vaccine candidates against *Neisseria*, and this approach can be used to design a universal vaccine against a wide range of pathogens. The epitope mapping pipeline we developed, which identifies B cell and helper T cell epitopes, can be used to discover peptide epitopes from the surface-exposed ClubSub-P clusters of other organisms as well. The C-terminal insertion motifs we obtained in the study will help the experimental scientists to ensure that the C-terminal insertion signals of the OMPs will be recognized by the BAMs of different organisms before they start their laborious experiments.

XI. Reference

1. Janeway CA, Travers P, Walport M, Shlomchik MJ: *Immunobiology: the immune system in health and disease.* Current Biology; 2005.
2. Peterson JW (Ed.). **Bacterial Pathogenesis**. Galveston, Texas: University of Texas Medical Branch; 1996.
3. Wilson J, Schurr M, LeBlanc C, Ramamurthy R, Buchanan K, Nickerson C: **Mechanisms of bacterial pathogenicity.** *Postgraduate medical journal* 2002, **78:**216-224.
4. Ratledge C, Dover LG: **Iron metabolism in pathogenic bacteria.** *Annual review of microbiology* 2000, **54:**881-941.
5. Purcell AW, McCluskey J, Rossjohn J: **More than one reason to rethink the use of peptides in vaccine design.** *Nature reviews Drug discovery* 2007, **6:**404-414.
6. Siegrist C-A (Ed.). **Vaccine immunology**, 5 edition: Saunders Elsevier; 2008.
7. Waterer GW, Wunderink RG: **Increasing threat of Gram-negative bacteria.** *Crit Care Med* 2001, **29:**N75-81.
8. Telford JL: **Bacterial genome variability and its impact on vaccine design.** *Cell Host Microbe* 2008, **3:**408-416.
9. Scarselli M, Giuliani MM, Adu-Bobie J, Pizza M, Rappuoli R: **The impact of genomics on vaccine design.** *Trends Biotechnol* 2005, **23:**84-91.
10. Rappuoli R: **Reverse vaccinology, a genome-based approach to vaccine development.** *Vaccine* 2001, **19:**2688-2691.
11. Movahedi AR, Hampson DJ: **New ways to identify novel bacterial antigens for vaccine development.** *Vet Microbiol* 2008, **131:**1-13.
12. Rappuoli R: **Reverse vaccinology.** *Curr Opin Microbiol* 2000, **3:**445-450.
13. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, et al: **Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing.** *Science* 2000, **287:**1816-1820.
14. Mora M, Donati C, Medini D, Covacci A, Rappuoli R: **Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach.** *Current Opinion in Microbiology* 2006, **9:**532-536.
15. Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, Tettelin H, Brettoni C, Iacobini ET, Rosini R, et al: **Identification of a universal Group B streptococcus vaccine by multiple genome screen.** *Science* 2005, **309:**148-150.
16. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".** *Proc Natl Acad Sci U S A* 2005, **102:**13950-13955.
17. Cassone A, Rappuoli R: **Universal Vaccines: Shifting to One for Many.** *mBio* 2010, **1:**e00042--00010.
18. Davies MN, Flower DR: **Harnessing bioinformatics to discover new vaccines.** *Drug Discovery Today* 2007, **12:**389-395.

19. Robert V, Volokhina EB, Senf F, Bos MP, Van Gelder P, Tommassen J: **Assembly factor Omp85 recognizes its outer membrane protein substrates by a species-specific C-terminal motif.** *PLoS Biology* 2006, **4:**e377.
20. Jungblut PR: **Proteome analysis of bacterial pathogens.** *Microbes and Infection* 2001, **3:**831-840.
21. Thein M, Sauer G, Paramasivam N, Grin I, Linke D: **Efficient Subfractionation of Gram-Negative Bacteria for Proteomics Studies.** *Journal of Proteome Research* 2010, **9:**6135-6147.
22. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FSL: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26:**1608-1615.
23. Berven FS, Flikka K, Jensen HB, Eidhammer I: **BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria.** *Nucleic Acids Research* 2004, **32:**W394--W399.
24. Remmert M, Linke D, Lupas AN, Söding J: **HHomp--prediction and classification of outer membrane proteins.** *Nucleic Acids Research* 2009, **37:**W446-451.
25. Desvaux M, Hébraud M, Talon R, Henderson IR: **Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue.** *Trends in microbiology* 2009, **17:**139-145.
26. Luirink J, von Heijne G, Houben E, de Gier J-W: **Biogenesis of inner membrane proteins in Escherichia coli.** *Annual review of microbiology* 2005, **59:**329-355.
27. Bendtsen J, Nielsen H, Widdick D, Palmer T, Brunak S: **Prediction of twin-arginine signal peptides.** *BMC Bioinformatics* 2005, **6:**167.
28. Tokuda H, Matsuyama S-i: **Sorting of lipoproteins to the outer membrane in E. coli.** *Biochimica et Biophysica Acta {(BBA)} - Molecular Cell Research* 2004, **1693:**5-13.
29. Kim KH, Aulakh S, Paetzel M: **The bacterial outer membrane β-barrel assembly machinery.** *Protein Science* 2012.
30. Arnold R, Brandmaier S, Kleine F, Tischler P, Heinz E, Behrens S, Niinikoski A, Mewes H-W, Horn M, Rattei T: **Sequence-based prediction of type III secreted proteins.** *PLoS Pathogen* 2009, **5:**e1000376.
31. Löwer M, Schneider G: **Prediction of Type III Secretion Signals in Genomes of Gram-Negative Bacteria.** *PLoS ONE* 2009, **4:**e5917.
32. Shen YQ, Burger G: **'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools.** *BMC bioinformatics* 2007, **8:**420.
33. Horler RSP, Butcher A, Papangelopoulos N, Ashton PD, Thomas GH: **EchoLOCATION: an in silico analysis of the subcellular locations of Escherichia coli proteins and comparison with experimentally derived locations.** *Bioinformatics* 2009, **25:**163-166.
34. Giombini E, Orsini M, Carrabino D, Tramontano A: **An automatic method for identifying surface proteins in bacteria: {SLEP}.** *BMC Bioinformatics* 2010, **11:**39.

35. Goudenège D, Avner S, Lucchetti-Miganeh C, Barloy-Hubler F: **CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources.** *BMC microbiology* 2010, **10:**88.
36. Overbeek R, Bartels D, Vonstein V, Meyer F: **Annotation of bacterial and archaeal genomes: improving accuracy and consistency.** *Chemical reviews* 2007, **107:**3431-3447.
37. Paramasivam N, Linke D: **ClubSub-P: Cluster-based subcellular localization prediction for Gram-negative bacteria and Archaea.** *Frontiers in Microbiology* 2011, **2:**1-14.
38. Juncker AS, Willenbrock H, von Heijne G, Brunak Sr, Nielsen H, Krogh A: **Prediction of lipoprotein signal peptides in Gram-negative bacteria.** *Protein Science* 2003, **12:**1652-1662.
39. Rose RW, Bruser T, Kissinger JC, Pohlschroder M: **Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway.** *Molecular Microbiology* 2002, **45:**943-950.
40. Dyrløv Bendtsen J, Nielsen H, von Heijne G, Brunak S: **Improved Prediction of Signal Peptides: SignalP 3.0.** *Journal of Molecular Biology* 2004, **340:**783-795.
41. Hiller K, Grote A, Scheer M, Münch R, Jahn D: **PrediSi: prediction of signal peptides and their cleavage positions.** *Nucleic acids research* 2004, **32:**W375-379.
42. Dariusz Plewczynskia LS, Adrian Tkaczb, Laszlo Kajanb, Liisa Holmc, Krzysztof Ginalskia And Leszek Rychlewskib: **The RPSP: Web server for prediction of signal peptides.** *Polymer* 2007, **48:**5493-5496.
43. Kall L, Krogh A, Sonnhammer ELL: **Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server.** *Nucl Acids Res* 2007, **35:**W429-432.
44. Imam S, Chen Z, Roos DS, Pohlschroder M: **Identification of surprisingly diverse type IV pili, across a broad range of gram-positive bacteria.** *PLoS One* 2011, **6:**e28919.
45. Tusnady GE, Simon I: **The HMMTOP transmembrane topology prediction server.** *Bioinformatics* 2001, **17:**849-850.
46. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL: **Predicting transmembrane protein topology with a hidden markov model: application to complete genomes.** *Journal of Molecular Biology* 2001, **305:**567-580.
47. Yu C-S, Chen Y-C, Lu C-H, Hwang J-K: **Prediction of protein subcellular localization.** *Proteins* 2006, **64:**643-651.
48. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.** *Bioinformatics (Oxford, England)* 2006, **22:**1658-1659.
49. Bagos PG, Tsirigos KD, Plessas SK, Liakopoulos TD, Hamodrakas SJ: **Prediction of signal peptides in archaea.** *Protein engineering, design & selection : PEDS* 2009, **22:**27-35.
50. Szabó Z, Stahl AO, Albers S-V, Kissinger JC, Driessen AJM, Pohlschröder M: **Identification of diverse archaeal proteins with class III signal peptides cleaved by distinct archaeal prepilin peptidases.** *Journal of bacteriology* 2007, **189:**772-778.

51. Biegert A, Mayer C, Remmert M, Söding J, Lupas AN: **The MPI Bioinformatics Toolkit for protein sequence analysis.** *Nucleic acids research* 2006, **34:**W335-339.
52. Starmer J, Stomp A, Vouk M, Bitzer D: **Predicting Shine-Dalgarno sequence locations exposes genome annotation errors.** *PLoS Computational Biology* 2006, **2:**e57.
53. Pallejà A, Harrington ED, Bork P: **Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions?** *BMC genomics* 2008, **9:**335.
54. Lukashin AV, Borodovsky M: **GeneMark.hmm: new solutions for gene finding.** *Nucleic Acids Res* 1998, **26:**1107-1115.
55. Delcher AL, Bratke KA, Powers EC, Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer.** *Bioinformatics* 2007**:**673-679.
56. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification.** *BMC Bioinformatics* 2010, **11:**119.
57. Knowles TJ, Scott-Tucker A, Overduin M, Henderson IR: **Membrane protein architects: the role of the BAM complex in outer membrane protein assembly.** *Nature Reviews Microbiology* 2009, **7:**206-214.
58. Bos MP, Robert V, Tommassen J: **Biogenesis of the gram-negative bacterial outer membrane.** *Annual Review of microbiology* 2007, **61:**191-214.
59. Sklar JG, Wu T, Kahne D, Silhavy TJ: **Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in Escherichia coli.** *Genes & Development* 2007, **21:**2473-2484.
60. Anwari K, Webb CT, Poggio S, Perry AJ, Belousoff M, Celik N, Ramm G, Lovering A, Sockett RE, Smit J, et al: **The evolution of new lipoprotein subunits of the bacterial outer membrane BAM complex.** *Molecular Microbiology* 2012, **84:**832-844.
61. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B: **Predicting transmembrane beta-barrels in proteomes.** *Nucleic Acids Research* 2004, **32:**2566-2577.
62. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *Journal of Molecular Biology* 1999, **292:**195-202.
63. Frith MC, Saunders NFW, Kobe B, Bailey TL: **Discovering Sequence Motifs with Arbitrary Insertions and Deletions.** *PLoS Computational Biology* 2008, **4:**e1000071.
64. Paramasivam N, Habeck M, Linke D: **Is the C-terminal insertional signal in Gram-negative bacterial outer membrane proteins species-specific or not?** *BMC Genomics (Under Revision)* 2012.
65. Suzuki R, Shimodaira H: **Pvclust: an R package for assessing the uncertainty in hierarchical clustering.** *Bioinformatics* 2006, **22:**1540-1542.
66. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S: **New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids.** *Journal of Medicinal Chemistry* 1998, **41:**2481-2491.

67. Lehr U, Schütz M, Oberhettinger P, Ruiz-Perez F, Donald JW, Palmer T, Linke D, Henderson IR, Autenrieth IB: **C-terminal amino acid residues of the trimeric autotransporter adhesin YadA of Yersinia enterocolitica are decisive for its recognition and assembly by BamA.** *Molecular Microbiology* 2010, **78:**932-946.
68. Fischer W, Schwan D, Gerland E, Erlenfeld GE, Odenbreit S, Haas R: **A plasmid-based vector system for the cloning and expression of Helicobacter pylori genes encoding outer membrane proteins.** *Molecular and General Genetics MGG* 1999, **262:**501-507.
69. Caoili SEC: **Benchmarking B-cell epitope prediction for the design of peptide-based vaccines: problems and prospects.** *Journal of biomedicine & biotechnology* 2010, **2010:**1-14.
70. Blythe MJ, Flower DR: **Benchmarking B cell epitope prediction: underperformance of existing methods.** *Protein Sci* 2005, **14:**246-248.
71. Ponomarenko J, Van Regenmortel M (Eds.): **B-cell epitope prediction**. Hoboken, NJ: John Wiley; 2009.
72. Söllner J, Heinzel A, Summer G, Fechete R, Stipkovits L, Szathmary S, Mayer B: **Concept and application of a computational vaccinology workflow.** *Immunome research* 2010, **6 Suppl 2:**S7.
73. Haste Andersen P, Nielsen M, Lund O: **Prediction of residues in discontinuous B-cell epitopes using protein 3D structures.** *Protein Sci* 2006, **15:**2558-2567.
74. El-Manzalawy Y, Dobbs D, Honavar V: **Predicting linear B-cell epitopes using string kernels.** *J Mol Recognit* 2008, **21:**243-255.
75. Larsen JE, Lund O, Nielsen M: **Improved method for predicting linear B-cell epitopes.** *Immunome Res* 2006, **2:**2.
76. Toussaint NC, Maman Y, Kohlbacher O, Louzoun Y: **Universal peptide vaccines - optimal peptide vaccine design based on viral sequence conservation.** *Vaccine* 2011, **29:**8745-8753.
77. Zhang L, Udaka K, Mamitsuka H, Zhu S: **Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools.** *Brief Bioinform* 2012, **13:**350-364.
78. Sette A, Vitiello A, Reherman B, Fowler P, Nayersina R, Kast WM, Melief C, Oseroff C, Yuan L, Ruppert J: **The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes.** *The Journal of Immunology* 1994, **153:**5586-5592.
79. Nielsen M, Justesen S, Lund O, Lundegaard C, Buus S: **NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure.** *Immunome Res* 2010, **6:**9.
80. Nielsen M, Lund O: **NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction.** *BMC Bioinformatics* 2009, **10:**296.
81. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C: **A generic method for assignment of reliability scores applied to solvent accessibility predictions.** *BMC Struct Biol* 2009, **9:**51.

XII. Curriculum vitae of Nagarajan Paramasivam

Born on 30th of May 1985, in Madurai, India. Indian. Single.

Education:

July 2002 – April 2005	B.sc., in Zoology (specialization in Biotechnology) Department of Zoology, The American College, Madurai, India.
July 2005- April 2007	M.sc in Bioinformatics Department of Bioinformatics, Bharathiar University, Coimbatore, India. May – July 2006 Summer internship with Prof. R. Nayak, Department of Molecular and Cell biology, Indian Institute of Science, India. Insilco Approach for Predicting T cell Epitopes from Conserved Hypothetical proteins of *Mycobacterium tuberculosis* H37Rv. December 2006 – April 2007 Master thesis with Dr. M. K. Mathew, National centre for biological science, India. How a voltage gated potassium channel opens?

Doctorate:

November 2008 -	Bacterial cell envelope, Department of Protein Evolution, Max-Planck institute of Developmental Biology and Eberhard-Karls University, Tübingen. PhD thesis: Identification of surface structures common to Gram-negative bacteria that are suitable for vaccine development; Supervisor: Prof. Dr. Friedrich Götz and PD. Dr. Dirk Linke.

Publications:

- Paramasivam N, Linke D. 2011. ClubSub-P: Cluster-based subcellular localization prediction for Gram-negative bacteria and Archaea. Frontiers in Microbiology 2.
- Paramasivam N, Habeck M, Linke D. 2012. Is the C-terminal insertional signal in Gram-negative bacterial outer membrane proteins species-specific or not? BMC Genomics (Under Revision).
- Shameer K, Nagarajan P, Gaurav K, Sowdhamini R. 2009. 3 PFDB- A database of Best Representative PSSM Profiles(BRPs) of Protein Families generated using a novel data mining approach. BioData mining 2: 8-8.
- Thein M, Sauer G, Paramasivam N, Grin I, Linke D. 2010. Efficient Subfractionation of Gram-Negative Bacteria for Proteomics Studies. Journal of Proteome Research 9: 6135-6147
- Upadhyay SK, Nagarajan P, Mathew M. 2009. Potassium channel opening: a subtle two-step. The Journal of Physiology 587: 3851-3868.

XIII. Published/Under revision research articles

Thein M, Sauer G, Paramasivam N, Grin I, Linke D

2010

Efficient Subfractionation of Gram-Negative Bacteria for Proteomics Studies. Journal of Proteome Research 9: 6135-6147

Pages 52 - 64

Paramasivam N, Linke D

2011

ClubSub-P: Cluster-based subcellular localization prediction for Gram-negative bacteria and Archaea. Frontiers in Microbiology 2

Pages 65 - 78

Paramasivam N, Habeck M, Linke D

2012

Is the C-terminal insertional signal in Gram-negative bacterial outer membrane proteins species-specific or not? BMC Genomics (Under Revision).

Pages 79 - 115

research articles Journal of proteome research

Efficient Subfractionation of Gram-Negative Bacteria for Proteomics Studies

Marcus Thein,[†,‡] Guido Sauer,[†,§] Nagarajan Paramasivam,[‡] Iwan Grin,[‡] and Dirk Linke*[,‡]

Department I, Protein Evolution and Department II, Biochemistry, Max Planck Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany

Received March 17, 2010

Proteomics studies of pathogenic bacteria are an important basis for biomarker discovery and for the development of antimicrobial drugs and vaccines. Especially where vaccines are concerned, it is of great interest to explore which bacterial factors are exposed on the bacterial cell surface and thus can be directly accessed by the immune system. One crucial step in proteomics studies of bacteria is an efficient subfractionation of their cellular compartments. We set out to compare and improve different protocols for the fractionation of proteins from Gram-negative bacteria into outer membrane, cytoplasmic membrane, periplasmic, and cytosolic fractions, with a focus on the outer membrane. Overall, five methods were compared, three methods for the fast isolation of outer membrane proteins and two methods for the fractionation of each cellular compartment, using *Escherichia coli* BL21 as a model organism. Proteins from the different fractions were prepared for further mass spectrometric analysis by SDS gel electrophoresis and consecutive in-gel tryptic digestion. Most published subfractionation protocols were not explicitly developed for proteomics applications. Thus, we evaluated not only the separation quality of the five methods but also the suitability of the samples for mass spectrometric analysis. We could obtain high quality mass spectrometry data from one-dimensional SDS-PAGE, which greatly reduces experimental time and sample amount compared to two-dimensional electrophoresis methods. We then applied the most specific fractionation technique to different Gram-negative pathogens, showing that it is efficient in separating the subcellular proteomes independent of the species and that it is capable of producing high-quality proteomics data in electrospray ionization mass spectrometry.

Keywords: membrane proteins • cell fractionation • bacterial pathogen • outer membrane

Introduction

Gram-negative bacteria all contain a complex cell envelope composed of the outer membrane (OM), the periplasm with the peptidoglycan layer, and the cytoplasmic membrane. Together with the cytoplasm, each of those subcellular compartments contains a unique set of proteins. It is possible today to predict proteomes and protein localization *in silico* based on genomic information, but the available computational methods are usually based on experimentally determined training sets, which are incomplete. Experimental verification of localization prediction is usually necessary; moreover, whether proteins are expressed under given growth conditions cannot be determined by bioinformatics (yet).

Proteomics studies aim at a large-scale characterization of *in vivo* localization, abundance, and post-translational modifications of proteins under varying growth conditions, and at identifying molecular interactions.[1] One major aim of proteomics research on pathogenic bacteria is the investigation of the bacterial cell envelope, and especially the OM, as this is the contact point of the bacterial cell to its environment, to host cells, and to the immune system. In fact, most virulence factors in pathogenic bacteria are localized on their cell surface.[2−4] To date, mass spectrometry-based proteomic databases of a number of Gram-negative bacteria are available, including pathogenic species such as *Escherichia coli*, *Vibrio cholerae*, *Pseudomonas aeruginosa*, *Haemophilus influenzae* or *Borrelia burgdorferi*.[4]

To undertake comprehensive proteomic studies of bacterial OMs, three main experimental steps have to be considered: (i) enrichment and purification of the OM; (ii) solubilization of outer membrane proteins (OMPs); and (iii) identification and characterization. Each of these three steps provides experimental challenges, often based on the intractable behavior of hydrophobic membrane proteins in the isolation of inner or outer membrane, in isoelectric focusing and in mass spectrometry experiments.[5−7] The first and most critical step is the enrichment and purification of OMs, where contaminations, for example, with highly abundant cytoplasmic proteins, must be avoided, or at least minimized.

* To whom correspondence should be addressed. Dirk Linke, Department I, Protein Evolution, Max Planck Institute for Developmental Biology, Spemannstrasse 35, 72076 Tübingen, Germany. Tel.: 49-7071-601-357. Fax: 49-7071-601-349. E-mail: dirk.linke@tuebingen.mpg.de.
† These authors contributed equally to this work.
‡ Department I, Protein Evolution.
§ Department II, Biochemistry.

10.1021/pr1002438

Cellular subfractions can be prepared based on the specific properties of proteins from different compartments. Isolation of the periplasm can be either achieved by the cold osmotic shock method[8] or by cell spheroblasting followed by differential centrifugation.[9] The cytoplasm can be isolated by cell lysis after separation of the periplasmic proteins and spinning down of the crude membranes. Inner and outer membranes of Gram-negative bacteria are usually prepared by lysozyme/EDTA lysis,[10] by French Press lysis[11] or by the use of commercially available lysis reagents.[12] These processes all yield membrane vesicles, which can be separated by density centrifugation using a sucrose gradient, as OM vesicles have a higher density than inner membrane (IM) vesicles.[13] Alternatively, inner and outer membranes can be separated by selective detergent treatment, using detergents which only dissolve the cytoplasmic membrane, such as Triton X-100,[9,11,14] or by washing of the vesicles with chaotropic agents such as sodium carbonate,[15,16] each followed by differential centrifugation.[17]

In this study, we examined the "somewhat confusing available literature"[17] of membrane proteome analysis and related literature of subcellular proteomics of Gram-negative bacteria and selected five different protocols for subcellular fractionation which we evaluated in detail. The main criteria apart from the specificity were the ease of use, the ability to yield samples that can directly be used in mass spectrometry experiments, and the applicability to different Gram-negative species. We evaluated the quality of the subfractions by immunoblotting, using antibodies raised against marker proteins of all cellular compartments. The efficiency of OM fractionation was then analyzed by liquid chromatography-coupled ESI-MS for all methods, comparing the results to the expected OM proteome predicted *in silico*. Finally, we show that the most efficient protocol can be adapted to other Gram-negative bacteria, namely to uropathogenic *Escherichia coli* 536, enteropathogenic *E. coli* 2348/69, the opportunistic pathogen *Pseudomonas aeruginosa*, and diarrhea-causing *Yersinia pseudotuberculosis*.

Materials and Methods

Bioinformatics. The whole proteome of *Escherichia coli* BL21 (NC_012947) was analyzed by PSORTb v.3.0.2,[18] HHOMP,[19] and BOMP[20] for outer membrane β-barrel proteins. The subcellular localization of the proteins identified in the mass spectrometry experiments was annotated by PSORTb v.3.0.2.[18]

Bacterial Strains and Growth Conditions. The strains used in this study were *Escherichia coli* BL21 harboring the plasmid pASK-IBA2, *Escherichia coli* 536, *Escherichia coli* 2348/69, *Pseudomonas aeruginosa* PAO1, and *Yersinia pseudotubercolusis* IP32953. Bacteria were grown in LB medium at 37 °C until cells were in the logarithmic growth phase at an OD_{600} of approximately 0.8. Cells were harvested by centrifugation.

Subfractionation Methods. Cellular subfractionation was performed according to five different methods, with methods 1, 2, and 3 focusing on the fast enrichment of OMs, and thus, of outer membrane proteins, and methods 4 and 5 focusing on the fractionation of proteins from cytoplasm, cytoplasmic membrane, periplasm and OM from one origin cell culture. Method 3 was developed in our laboratory in the course of this study.

Method 1 for OM separation is based on previously published protocols and uses lysozyme/EDTA lysis followed by selective detergent treatment.[9,14] In detail, 25 mL of cells were pelleted and resuspended in 500 μL 0.2 M Tris-HCl pH 8, 1 M sucrose, 1 mM EDTA. 100 μL of lysozyme (Sigma-Aldrich) (5 mg/mL in dH_2O) were added, vortexed and incubated for 5 min at RT. Two ml of dH_20 were added and incubated for 20 min at RT until spheroblast formation is observed under the microscope. Then 3 mL 50 mM Tris-HCl pH 8, 2% (w/v) Triton X-100, 10 mM $MgCl_2$ and 50 μL DNase I (Applichem) (1 mg/mL in dH_2O) were added and mixed until the suspension was clear. The mixture was ultracentrifuged at 85 000× *g* for 30 min at 4 °C. The pellet containing the OM was washed in 750 μL 50 mM Tris-HCl pH 8, 2% (w/v) Triton X-100, 10 mM $MgCl_2$, centrifuged at ≥85 000× *g* for 20 min at 4 °C, and was finally washed in 500 μL dH_2O three times and stored at −20 °C.

Method 2 for OM preparation is based on French press cell lysis followed by treatment with chaotropic reagents.[16] In detail, 50 mL of cells were pelleted and resuspended in 6 mL 0.1 M Tris-HCl pH 7.3 supplemented with 7 mg of DNase I. The cells were ruptured in a French Press with two passes at 10^8 Pa. Incompletely lysed cells were removed by centrifugation at 4000 rpm for 15 min (Eppendorf 5810 R centrifuge). The supernatant was diluted with ice-cold 0.1 M sodium carbonate pH 11 to a final volume of 60 mL and stirred 1 h at 4 °C. Then the suspension was ultracentrifuged at 120 000× *g* for 1 h at 4 °C. The pellet containing the OM was washed in 2 mL 0.1 M Tris-HCl pH 7.3, centrifuged at ≥85 000× *g* for 20 min at 4 °C, and was finally washed in 500 μL dH_2O three times and stored at −20 °C.

Method 3 combines outer membrane separation by selective detergent use (as used in method 1) with treatment with chaotropic reagents (as used in method 2). In detail, 25 mL of cells were pelleted and resuspended in 500 μL 0.2 M Tris-HCl pH 8, 1 M sucrose, 1 mM EDTA. One-hundred microliters of lysozyme (5 mg/mL in dH_2O) were added, vortexed and incubated for 5 min at RT. Two ml of dH_20 were added and incubated for 20 min at RT until spheroblast formation was observable under the microscope. Then 3 mL 50 mM Tris-HCl pH 8, 2% (w/v) Triton X-100, 10 mM $MgCl_2$ and 50 μL DNase I (1 mg/mL in dH_2O) were added and mixed until suspension became clear. The mixture was ultracentrifuged at 85 000× *g* for 30 min at 4 °C. The pellet containing the OM was washed in 750 μL 50 mM Tris-HCl pH 8, 2% (w/v) Triton X-100, 10 mM $MgCl_2$, centrifuged at ≥85 000× *g* for 20 min at 4 °C. The pellet containing the outer membrane was then resuspended in ice-cold 0.1 M sodium carbonate pH 11 to a final volume of 60 mL and stirred 1 h at 4 °C. Then the suspension was ultracentrifuged at 120 000× *g* for 1 h at 4 °C. The pellet containing the OM was washed in 2 mL 0.1 M Tris-HCl pH 7.3, centrifuged at ≥85 000× *g* for 20 min at 4 °C, and was finally washed in 500 μL dH_2O three times and stored at −20 °C.

Methods 4 and 5 are partly overlapping and are based on previously published protocols for subcellular fractionation.[9,13,14,21] 1 L of cells was pelleted and resuspended in 10 mL 0.2 M Tris-HCl pH 8, 1 M sucrose, 1 mM EDTA and lysozyme was added to a final concentration of 1 mg/mL. The suspension was mixed and incubated for 5 min at RT. 40 mL dH_2O were added to the swirling mixture before placing on ice (spheroblasting can be checked under microscope). Then cells were centrifuged at 200 000× *g* for 45 min at 4 °C. The supernatant contained the periplasmic fraction. The pellet was resuspended in 7.5 mL ice-cold 10 mM Tris-HCl pH 7.5, 5 mM EDTA, 0.2 mM DTT supplemented with 50 μL DNase (1 mg/mL). The cells were ruptured in a French Press with two passes at 10^8 Pa. Unbroken cells were spun down by centrifugation at 4000 rpm for 10 min at 4 °C (Eppendorf 5810 R centrifuge). Then the supernatant was centrifuged at 280 000−300 000× *g*

for 2−4 h at 4 °C. The supernatant contained the cytoplasmic fraction and the pellet contained the crude membranes.

The difference between method 4 and 5 lies in the process how inner and outer membranes are separated: in method 4, inner and outer membrane was separated by selected detergent treatment in a similar manner as used in method 1. In detail, the crude membrane pellet was resuspended in 9 mL 50 mM Tris-HCl pH 8, 2% (w/v) Triton X-100, 10 mM $MgCl_2$ and centrifuged at 85 000× *g* for 30 min at 4 °C. The supernatant contained the cytoplasmic membrane fraction. The pellet was washed in 1 mL 50 mM Tris-HCl pH 8, 2% (w/v) Triton X-100, 10 mM $MgCl_2$, centrifuged at ≥85 000× *g* for 20 min at 4 °C, and was finally washed in 500 μL dH_2O three times and stored at −20 °C.

Method 5 is based on the separation of inner and outer membrane vesicles according to their different densities.[13,21] In particular the crude membrane pellet was resuspended in 1 mL 10 mM Tris pH 7.5, 15% sucrose (w/w), 5 mM EDTA, and 0.2 mM DTT. The sucrose gradient was prepared by layering 2.25 mL each of 50, 45, 35 and 30% sucrose solutions (w/w) over a cushion of 1 mL 55% sucrose, all sucrose solutions contained 10 mM Tris-HCl pH 7.5, 5 mM EDTA. The membrane suspension was layered on the top of the gradient and centrifuged at 250 000× *g* in a swinging bucket rotor for 12−16 h at 4 °C. The visible bands were carefully collected with a coarse syringe needle. The lower density band (upper band) contained the cytoplasmic membrane and the higher density band (lower band) contained the outer membrane fraction; a third, mixed band between them was discarded. The membrane fractions were diluted 1:1 with dH_2O, centrifuged at ≥85 000× *g* for 20 min at 4 °C, and was finally washed in 500 μL dH_2O three times and stored at −20 °C.

SDS-PAGE and Immunoblotting. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was performed using the Laemmli gel system.[22] In detail, proteins were boiled for 5 min in 4× SDS sample buffer before loading the gel and then separated by 8−18% SDS-PAGE under denaturing conditions. The gels were usually silver-stained,[23] only gels used for mass spectrometric analysis were stained with colloidal Coomassie.[24] After solubilization of the inner and outer membrane fractions in SDS sample buffer and separation by electrophoresis, inner and outer membrane proteins are mostly free from lipids and other membrane components and can easily be extracted from the gels for further mass spectrometric analysis as described below.

Immunoblotting was performed with a standard semi dry protocol[25] using 0.45 μm nitrocellulose membranes for protein immobilization.[26] Bound antibodies were visualized using alkaline phosphatase conjugated antirabbit or antimouse antibodies (Jackson ImmunoResearch) and 5-bromo-4-chloro-3-indolylphosphate/nitroblue tetrazolium.

Mass Spectrometric Analysis. After staining with colloidal Coomassie whole lanes from one-dimensional SDS gels were cut into 12 bands which were subsequently in-gel digested with trypsin (Promega) in principle as published by Shevchenko et al.[27] The doubly extracted peptides were desalted with self-made microcolumns[28] and separated by reversed-phase HPLC (nanoLC2D, Eksigent) using a fused silica column of 14 cm length, 75 μm in diameter and 8 μm tip opening (PicoTip, New Objective) packed in-house with ReproSil-Pur C18-AQ, 3 μm (Dr. Maisch GmbH). A stepwise gradient from 3% to 80% buffer B with buffer A (0.1% formic acid in water) and buffer B (0.1% formic acid in acetonitrile) was applied over a run time of 40 min at a flow rate of 160 nL/min. Mass spectrometric analysis was performed using an ion trap (HCTultra PTM Discovery, Bruker Daltonics) equipped with a nanoESI source (Proxeon Biosystems).

Data-dependent acquisition was performed in positive-ion mode where 3−6 scans were added for each precursor ion scan in Standard Enhanced scan mode with a scan range from 300 to 1200 *m*/*z*. Up to 5 precursors were isolated and fragmented in UltraScan mode following each parent ion scan with active exclusion of 20 s. Mascot generic data files were created using DataAnalysis V4.0 SP1 (Bruker Daltonics) with the following settings: signal was considered as peaks when showing a signal-to-noise ratio better than 5 and exceeded both 10% area and intensity thresholds. Peak detection algorithm V2.0 was used, while no background subtraction was applied. Intensity threshold for AutoMS(n) detection was set to 5000 and up to 3000 fragmentation spectra were allowed. Deconvolution parameters were set to auto for MS and MS/MS data. Resulting peak lists of each sample were combined using Mascot Daemon and searched against the Swiss-Prot (V57.11) or NCBI_nr (as on 12/13/2009) protein database using Mascot Server (V2.2). The following settings were used: digestion with trypsin allowing 1 miss cleavage, carbamylation of cysteines as fixed modification, oxidation of methionine as variable, 0.3 Da peptide mass accuracy and 0.3 Da for fragmentation masses. MudPiT scoring was used. The threshold score for identified peptides was set to 10. Mascot protein hits matching to identical entries of other organisms/strains than the used ones were assigned to find the respective gene identifier of the correct organism/strain using the BLASTp algorithm against the predicted proteome of the respective organism.[29] Hits resulting in e-values greater than 10^{-3} were rejected.

Table 1. Predicted Outer Membrane Proteome of *E. coli* BL21

prediction	β-barrels	β-barrels and lipoproteins	reference
PSORTb v.3.0.2	N/A	83	18
HHOMP	69	N/A	19
BOMP	73	N/A	20

Results and Discussion

Predicted Outer Membrane Proteome. Integral membrane proteins from the cytoplasmic and outer membrane differ in secondary structure content and amino acid composition. Cytoplasmic membrane proteins are mainly composed of hydrophobic α-helices, whereas outer membrane proteins (OMPs) are mainly comprised of β-barrel motifs.[30−33] In addition, many lipoproteins are found on the cytoplasmic and outer membrane.[17,34,35] To obtain an estimate of the theoretical number of OMPs, we used different algorithms to predict OMPs of *E. coli* BL21 (Table 1). (a) PSORTb v.3.0.2 predicts 5 different subcellular localizations for Gram-negative bacteria, but does not distinguish between β-barrel proteins and lipoproteins in its OMP prediction;[18] 83 OMPs were identified using this algorithm. (b) HHOMP identifies β-barrel OMPs by detecting their homologous relationships to known OMPs;[19] 69 OMPs were found using this software. (c) BOMP was developed to predict integral β-barrel OMPs and uses C-terminal patterns and amino acid compositional information of OMPs with known structure;[20] it predicts 73 OMPs. Altogether, 121 different OM proteins were predicted, 38 of them were consistently predicted by all three tools. The Venn diagram in Figure 1 shows the overlap and the uniquely predicted OMPs, and

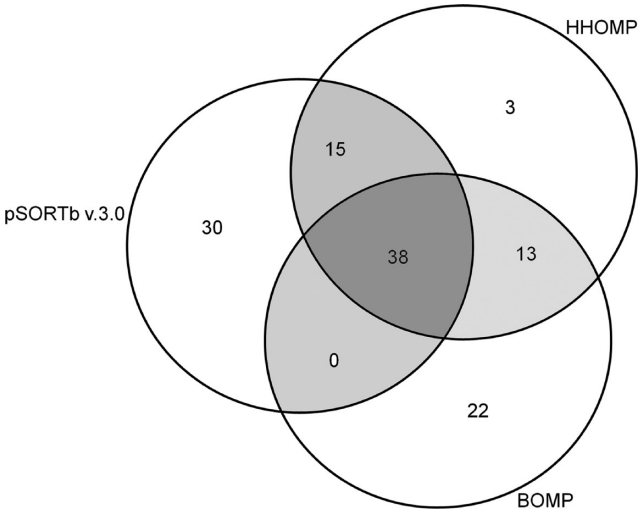


Figure 1. Venn diagram depicting the overlapping and uniquely predicted OM proteins in *E. coli* BL21. Altogether, 121 different OM proteins were predicted, 83 of those by PSORTb v.3.0.2,[18] 69 by HHOMP[19] and 73 by BOMP.[20]

clearly displays the need for experimental validation of *in silico* localization predictions.

Outer Membrane Isolation from *E. coli* BL21 Cells. We established five different subfractionation methods based on literature,[9,13,14,16,21,36] with the aim of evaluating their ease and speed of use, their quality, and their usefulness for mass spectrometry experiments. A flow-chart representation of the five methods is shown in Figure 2. Methods 1, 2, and 3 represent fast and easy protocols for OM isolation; method 1 is based on lysozyme/EDTA lysis followed by selective detergent treatment and method 2 is based on French Press lysis followed by chaotropic reagents treatment, respectively. In method 3, we combined the techniques used in method 1 and 2. Methods 4 and 5 are more extensive and facilitate the isolation of all subcellular compartments from one bacterial cell culture. They are based on the techniques of lysozyme/EDTA lysis followed by selective detergent treatment or sucrose density centrifugation, respectively. To compare the effectiveness of the five different subfractionation protocols, outer membrane fractions (OMFs) of the laboratory strain *E. coli* BL21 were prepared. The obtained OMFs contained numerous proteins with molecular weights ranging from approximately 10 kDa to above 130 kDa as shown in an 8−18% SDS-PAGE (Figure 3). The most prominent protein bands within all obtained OMFs display molecular weights between approximately 15 and 36 kDa. Those molecular weights correspond to well-known predominant *E. coli* OMPs, including among others OmpA, OmpF and OmpX.[16,37] The presence of OmpX is also shown by immunoblotting in Figure 5B. Apart from these eminent protein bands, there are also obvious variations in the band pattern of fractions prepared with different methods, e.g. at a molecular weight above approximately 50 kDa. The band pattern similarity between method 1 and 4 OMFs is remarkably high, which is not a surprise considering that both methods are based on selective detergent use.

Challenges of Outer Membrane Isolation with Regard to Proteomics Studies. One major challenge is the minimization of contaminating proteins from other cellular compartments. Even though such contaminations can be reduced by OM enrichment, a certain number of non-OM proteins in OMFs is unavoidable. This is partly due to the distinct protein properties on which the fractionation is based (mostly density and hydrophobicity). For example, several studies show that abundant, hydrophobic ribosomal proteins represent a major contaminant in OMFs.[36,38,39] When OMFs have been successfully enriched, one still faces the challenge of effectively solubilizing OMPs in order to perform further proteomic analysis. Many groups tried to overcome those limitations by using a large variety of solubilizing agents and variations in the electrophoresis technology.[40−43] In contrast to most published OM proteomics studies that rely on two-dimensional gel techniques, we decided to separate the obtained OMFs in one-dimensional (1D) SDS-PAGE as it is a fast, robust, and cheap way for a crude separation of proteins. Both techniques have limitations in their utility for fractionation of membrane proteins, but by using 1D SDS-PAGE we were able to use the strong ionic detergent SDS for efficient solubilization of hydrophobic proteins, which would be detrimental to isoelectric focusing.[44] Protein separation with 1D SDS-PAGE does not allow us to precisely attribute an identified protein to a spot in the gel, but this procedure minimizes the risk of overlooking low abundance OMPs and requires a much lower amount of sample compared to typical 2D gel approaches. In most 2D gel approaches, only a limited number of strongly hydrophobic membrane proteins have been detected.[7,16,45] 2D gels poorly resolve basic proteins and hydrophobic proteins, especially those with more than three transmembrane-spanning regions,[46] which generally have alkaline p*I*s and are poorly soluble in the aqueous media used for isoelectric focusing.

Mass Spectrometric Analysis of Outer Membrane Fractions. To evaluate the five different fractionation methods in terms of specificity and compatibility with electrospray ionization (ESI) mass spectrometry SDS-PAGE was performed as described, and the resulting gel lanes were cut into 12 slices each. The bands were subjected to in-gel tryptic digestion, and extracted peptides were then analyzed in liquid chromatography-coupled tandem mass spectrometry. To increase the significance of the comparative method analysis we run each experiment three times and then assessed the reproducibility. For method 1 we found 103 proteins identified in all three replicates, 67 in two and 86 to be present only in one of 256 proteins identified overall. For method 2, we found 71, 94, and 113 of 278 proteins, respectively. Method 3 yielded 59, 81, and 273 of 413 proteins, method 4 in 229, 109, and 94 of 432, and method 5 resulted in 64, 75, and 195 of 334 proteins. All proteins identified are listed in Table S1 (Supporting Information) including scores, sequence coverage and emPAI value. We want to emphasize that the identification of proteins with a high score was highly reproducible. For example, 49 of the top 50 scoring proteins of method 1 were found in all three replicates. In contrast, many low-scoring (i.e., low abundance) proteins were identified in one or two experiments only.

The localization of the identified proteins was annotated using PSORTb v.3.0.2 predictions. On the basis of these predictions, we evaluated the efficiency of OM separation (given by the number of non-OMPs in the sample) and the ability of the methods to identify as many OMPs as possible in a single experiment.

Similar to previous *E. coli* OM proteome studies[47,48] we found a variety of proteins with other cellular localization than OM in all OMFs, mostly cytoplasmic proteins. Depending on the method, we identified 256 (for method 1) to 432 (for method

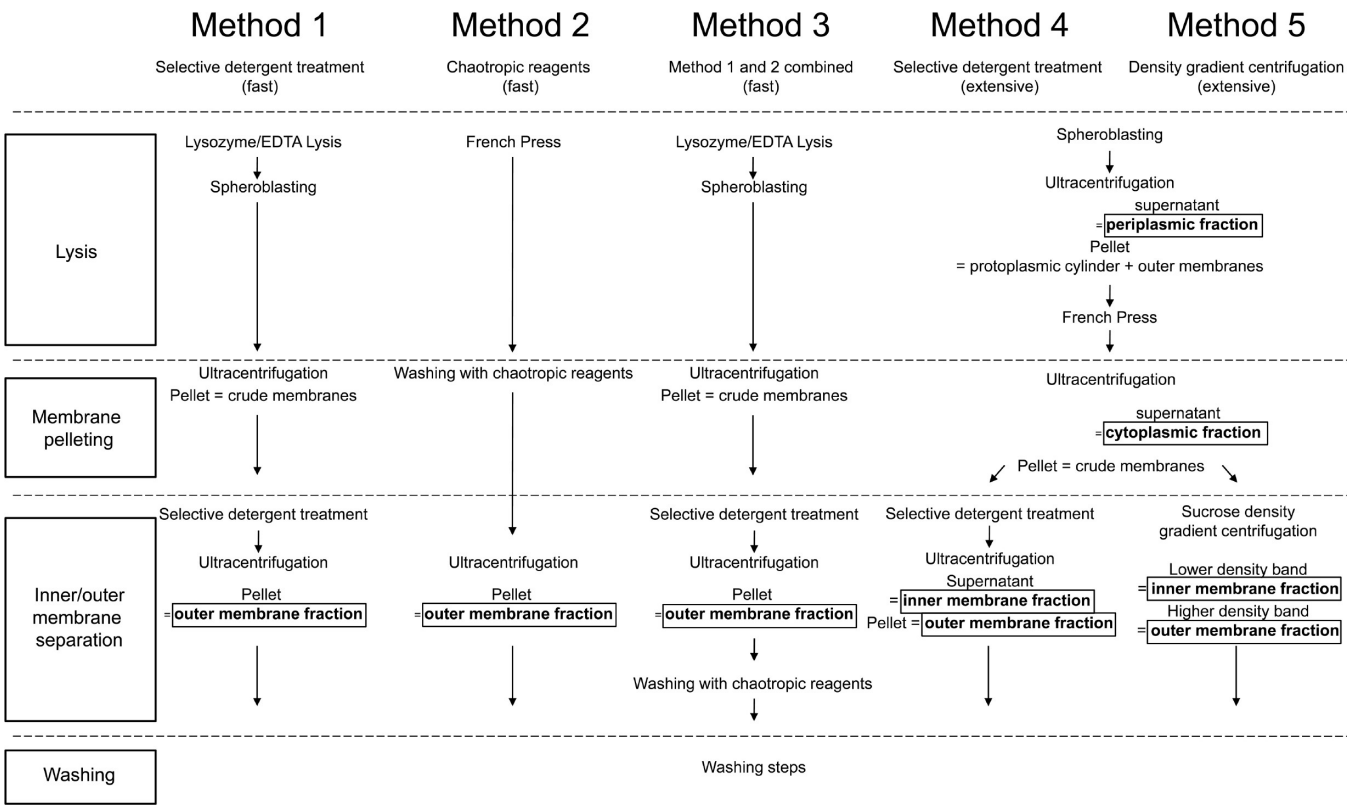


Figure 2. Flow-chart of the different subfractionation methods. Methods 1, 2, and 3 can be used for the fast enrichment of outer membrane proteins, and methods 4 and 5 facilitate the separation of all subcellular fractions from one origin cell culture.

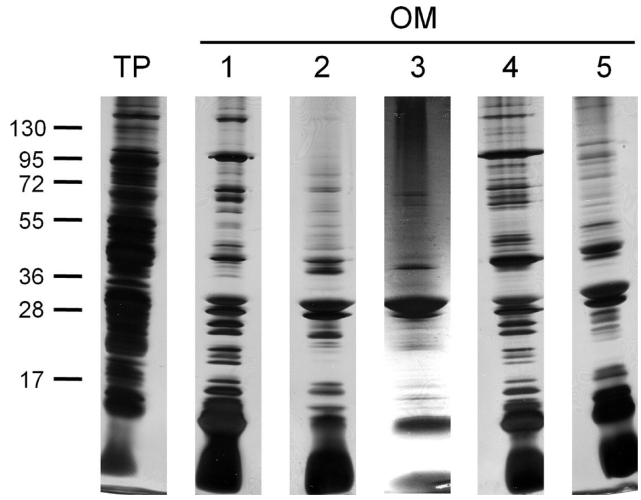


Figure 3. SDS-PAGE analysis of outer membrane fractions of *E. coli* BL21 prepared with methods 1−5. Approximately 10 μg of total protein (TP) and outer membrane (OM) fractions were separated by 8−18% SDS-PAGE. The positions of molecular mass standards (in kDa) are shown on the left of the silver-stained gel.

4) proteins in the OMFs. From this number approximately one-fourth (method 1) to one-half (method 2 and 5) were membrane proteins. Approximately 21% (method 3) to 41% (method 1) of those membrane proteins were annotated as OMPs by PSORTb v.3.0.2. These distributions are comparable to those of other OM proteomic studies.[39,47−49] 63% of all identified proteins were cytoplasmic in method 1 OMFs, 37% in method 2 OMFs, 55% in method 3 OMFs, 57% in method 4 OMFs and 43% in method 5 OMFs. Up to 42% of the cytoplasmic proteins were ribosomal proteins (method 2). This finding agrees with other *E. coli* OM proteome studies which similarly detected ribosomal and other highly abundant soluble cytoplasmic compounds as major contaminants in OMFs.[17,39,48] Contamination of OMFs with periplasmic proteins was low, representing maximally 3% of the total number of identified proteins in method 2 and 5 OMFs, and in method 1, 3, and 4 OMFs periplasmic proteins were only 2% of all proteins. Cytoplasmic membrane protein contaminants ranged from 16% (in method 1 OMFs) to 37% (in method 2 OMFs). In addition, in all OMFs one or 2 proteins were predicted to be extracellular. Last but not least, a number of proteins with unknown cellular location (according to PSORTb v.3.0.2 predictions) were identified in all OMFs. These proteins were 8% of all identified proteins in method 1 OMFs, 10% in method 2 OMFs, 6% in method 3 OMFs, 8% in method 4 OMFs and 10% in method 5 OMFs. Note that the protein identification by mass spectrometry does not include any precise quantitation; from the band pattern and the intensity of some well-known OMP bands in Figure 3, one can assume that a high percentage of the total protein in the OMFs are OMPs, but that only a few OMPs are present in high quantity.

Comparative Outer Membrane Proteomics of the Five Fractionation Methods. Altogether, 44 different OMPs have been unambiguously identified using the five separation methods (Table 2). The number of identified OMPs is comparable to the number of identified proteins in previous *E. coli* OM proteome studies.[16,39,48] In detail, 28 OMPs have been identified using method 1, 36 OMPs using method 2, 32 OMPs using method 3, 37 OMPs using method 4 and 37 OMPs using method 5. Twenty-one of all identified OMPs were identically found with all five OM separation methods. One outer membrane protein was exclusively found in method 1 OMFs, two

Table 2. Outer Membrane Proteome of *E. coli* BL21 Prepared According to Methods 1−5[a]

method 1		method 2		method 3		method 4		method 5	
gi number	gene description	gi number	gene description	gi number	gene description	gi number	gene description	gi number	gene description
gil253774014	OmpA domain protein transmembrane region-containing protein	gil253773366	LPP repeat-containing protein	gil253774042	porin Gram-negative type	gil253774014	OmpA domain protein transmembrane region-containing protein	gil253773366	LPP repeat-containing protein
gil253774042	porin Gram-negative type	gil253774042	porin Gram-negative type	gil253774014	OmpA domain protein transmembrane region-containing protein	gil253774042	porin Gram-negative type	gil253774042	porin Gram-negative type
gil253773366	LPP repeat-containing protein	gil253774014	OmpA domain protein transmembrane region-containing protein	gil253773366	LPP repeat-containing protein	gil253773366	LPP repeat-containing protein	gil253775338	porin LamB type
gil253775338	porin LamB type	gil253775338	porin LamB type	gil253775338	porin LamB type	gil253775338	porin LamB type	gil253774796	outer membrane protein assembly complex, YaeT protein
gil253772127	type I secretion outer membrane protein, TolC family	gil253772127	type I secretion outer membrane protein, TolC family	gil253772127	type I secretion outer membrane protein, TolC family	gil253772127	type I secretion outer membrane protein, TolC family	gil253774014	OmpA domain protein transmembrane region-containing protein
gil253774914	Organic solvent tolerance protein	gil253774914	Organic solvent tolerance protein	gil253774183	virulence-related outer membrane protein	gil253774288	peptidoglycan-associated lipoprotein	gil253774914	Organic solvent tolerance protein
gil253774796	outer membrane protein assocembly complex, YaeT protein	gil253774796	outer membrane protein assocembly complex, YaeT protein	gil253774288	peptidoglycan-assocociated lipoprotein	gil253774914	Organic solvent tolerance protein	gil253774183	virulence-related outer membrane protein
gil253774183	virulence-related outer membrane protein	gil253774288	peptidoglycan-assocociated lipoprotein	gil253774914	Organic solvent tolerance protein	gil253774796	outer membrane protein assembly complex, YaeT protein	gil253774192	TonB-dependent siderophore receptor
gil253774192	TonB-dependent siderophore receptor	gil253774183	virulence-related outer membrane protein	gil253774796	outer membrane protein assembly complex, YaeT protein	gil253774183	virulence-related outer membrane protein	gil253774288	peptidoglycan-assocociated lipoprotein
gil253773400	17 kDa surface antigen	gil253774192	TonB-dependent siderophore receptor	gil253773400	17 kDa surface antigen	gil253774192	TonB-dependent siderophore receptor	gil253773400	17 kDa surface antigen
gil253774288	peptidoglycan-assocociated lipoprotein	gil253774433	TonB-dependent siderophore receptor	gil253771615	OmpA/MotB domain protein	gil253773400	17 kDa surface antigen	gil253772127	type I secretion outer membrane protein, TolC family
gil253772594	outer membrane assembly lipoprotein YfgL	gil253773400	17 kDa surface antigen	gil253772733	membrane protein involved in aromatic hydrocarbon degradation	gil253772509	outer membrane assembly lipoprotein YfiO	gil253772594	outer membrane assembly lipoprotein YfgL
gil253771615	OmpA/MotB domain protein	gil253771615	OmpA/MotB domain protein	gil253772594	outer membrane assembly lipoprotein YfgL	gil253771615	OmpA/MotB domain protein	gil253772733	membrane protein involved in aromatic hydrocarbon degradation
gil253772733	membrane protein involved in aromatic hydrocarbon degradation	gil253772733	membrane protein involved in aromatic hydrocarbon degradation	gil253774192	TonB-dependent siderophore receptor	gil253774373	Rare lipoprotein B	gil253771615	OmpA/MotB domain protein
gil253774373	Rare lipoprotein B	gil253772594	outer membrane assembly lipoprotein YfgL	gil253773263	MltA-interacting MipA family protein	gil253773263	MltA-interacting MipA family protein	gil253774433	TonB-dependent siderophore receptor
gil253772509	outer membrane assembly lipoprotein YfiO	gil253772509	outer membrane assembly lipoprotein YfiO	gil253774373	Rare lipoprotein B	gil253772594	outer membrane assembly lipoprotein YfgL	gil253772509	outer membrane assembly lipoprotein YfiO
gil253773987	polysaccharide export protein	gil253774776	outer membrane lipoprotein	gil253772631	NlpBDapX family lipoprotein	gil253774776	outer membrane lipoprotein	gil253774373	Rare lipoprotein B
gil253774776	outer membrane lipoprotein	gil253774373	Rare lipoprotein B	gil253774776	outer membrane lipoprotein	gil253774433	TonB-dependent siderophore receptor	gil253773574	TonB-dependent receptor

Table 2. Continued

method 1		method 2		method 3		method 4		method 5	
gi number	gene description	gi number	gene description	gi number	gene description	gi number	gene description	gi number	gene description
gi|253772631	NlpBDapX family lipoprotein	gi|253772402	Peptidase M23	gi|253773241	outer membrane lipoprotein, Slp family	gi|253772733	membrane protein involved in aromatic hydrocarbon degradation	gi|253772402	Peptidase M23
gi|253774604	nucleoside-specific channel-forming protein Tsx	gi|253773534	TonB-dependent receptor plug	gi|253774604	nucleoside-specific channel-forming protein Tsx	gi|253773987	polysaccharide export protein	gi|253775091	TonB-dependent siderophore receptor
gi|253772730	VacJ family lipoprotein	gi|253773263	MltA-interacting MipA family protein	gi|253772509	outer membrane assembly lipoprotein YfiO	gi|253774604	nucleoside-specific channel-forming protein Tsx	gi|253772631	NlpBDapX family lipoprotein
gi|253773263	MltA-interacting MipA family protein	gi|253773574	TonB-dependent receptor	gi|253772294	Peptidase M23	gi|253773241	outer membrane lipoprotein, Slp family	gi|253774604	nucleoside-specific channel-forming protein Tsx
gi|253773534	TonB-dependent receptor plug	gi|253775091	TonB-dependent siderophore receptor	gi|253773987	polysaccharide export protein	gi|253772631	NlpBDapX family lipoprotein	gi|253774776	outer membrane lipoprotein
gi|253773574	TonB-dependent receptor	gi|253772631	NlpBDapX family lipoprotein	gi|253772730	VacJ family lipoprotein	gi|253772294	Peptidase M23	gi|253773263	MltA-interacting MipA family protein
gi|253772330	MltA domain protein	gi|253774604	nucleoside-specific channel-forming protein Tsx	gi|253775091	TonB-dependent siderophore receptor	gi|253772402	Peptidase M23	gi|253772730	VacJ family lipoprotein
gi|253771544	Peptidase M23	gi|253775553	Phospholipase A(2)	gi|253772402	Peptidase M23	gi|253775091	TonB-dependent siderophore receptor	gi|253775151	surface antigen (D15)
gi|253772490	SmpA/OmlA domain protein	gi|253772730	VacJ family lipoprotein	gi|253773630	porin Gram-negative type	gi|253772730	VacJ family lipoprotein	gi|253774450	RND efflux system, outer membrane lipoprotein, NodT family
gi|253774450	RND efflux system, outer membrane lipoprotein, NodT family	gi|253773897	flagellar L-ring protein	gi|253775151	surface antigen (D15)	gi|253772490	SmpA/OmlA domain protein	gi|253773987	polysaccharide export protein
		gi|253774280	Pectinesterase	gi|253773534	TonB-dependent receptor plug	gi|253775221	Lipocalin family protein	gi|253772330	MltA domain protein
		gi|253774780	copper resistance lipoprotein NlpE	gi|253774433	TonB-dependent siderophore receptor	gi|253775553	Phospholipase A(2)	gi|253772294	Peptidase M23
		gi|253772330	MltA domain protein	gi|253772330	MltA domain protein	gi|253775151	surface antigen (D15)	gi|253773534	TonB-dependent receptor plug
		gi|253774450	RND efflux system, outer membrane lipoprotein, NodT family	gi|253775553	Phospholipase A(2)	gi|253773897	flagellar L-ring protein	gi|253775553	Phospholipase A(2)
		gi|253773241	outer membrane lipoprotein, Slp family			gi|253774780	copper resistance lipoprotein NlpE	gi|253774780	copper resistance lipoprotein NlpE
		gi|253772294	Peptidase M23			gi|253774280	Pectinesterase	gi|253774280	Pectinesterase
		gi|253774732	porin Gram-negative type			gi|253773574	TonB-dependent receptor	gi|253772490	SmpA/OmlA domain protein
		gi|253772919	TonB-dependent receptor plug			gi|253771665	outer membrane lipoprotein, Slp family	gi|253773897	flagellar L-ring protein
						gi|253773746	OmpW family protein	gi|253773241	outer membrane lipoprotein, Slp family

[a] This table includes all proteins identified by nanoLC−MS/MS that were annotated as outer membrane proteins by PSORTb v.3.0.2.

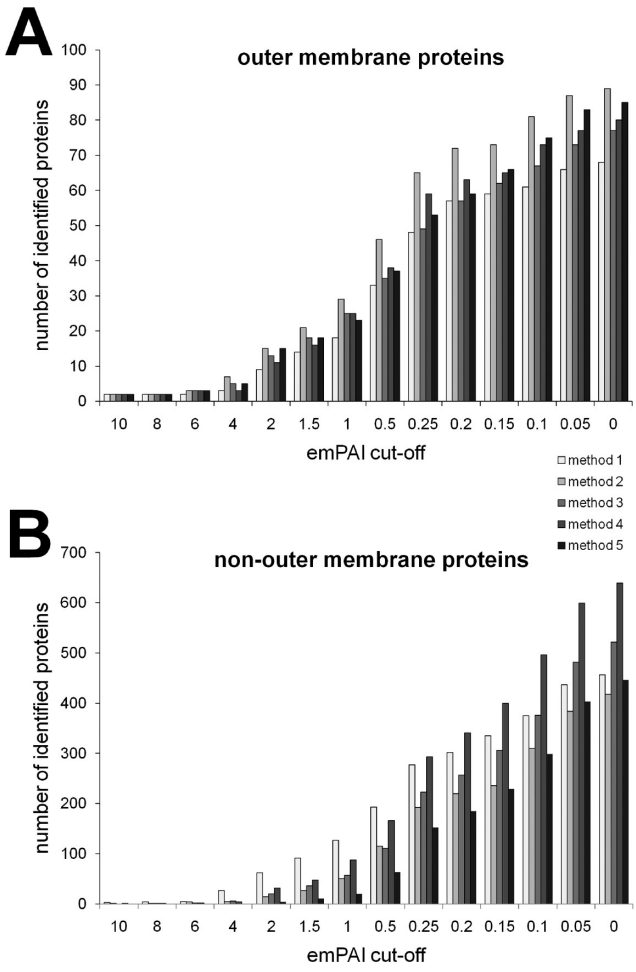


Figure 4. Number of identified proteins at defined emPAI cutoff levels (A). Outer membrane proteins and (B) nonouter membrane proteins according to PSORTb v.3.0.2 predictions were identified in ESI−MS experiments using OMFs prepared with methods 1−5. The abundance of OMPs and non-OMPs in all three replicates was compared at different cutoff levels of the emPAI (exponentially modified protein abundance index) of the single identified protein matches. The emPAI offers an approximate, relative quantitation of proteins in a mixture, where high emPAI scores denote a high abundance of the protein in the sample.[50] Note that above emPAI scores of 0.25, the number of non-OMPs continues to increase exponentially, while the number of found OMPs stagnates, suggesting that many non-OMP impurities are present only in low amounts.

proteins were exclusively found in method 2 OMFs, one protein was exclusively found in method 3 OMF, and three proteins were exclusively identified using method 4.

To obtain more quantitative information, we evaluated the emPAI scores (exponentially modified protein abundance index) of the identified proteins. The emPAI offers an approximate, relative quantitation of proteins in a mixture, where high emPAI scores denote a high abundance of the protein in the sample.[50] The number of identified OMPs and non-OMPs at defined emPAI cut-offs derived from all three replicates are displayed in Figure 4. In general, OMPs show significantly higher emPAI scores than proteins not localized in the OM. The histograms in figure 4A show that most of the OMPs were identified with an emPAI >0.25 independent of the method used for OM fractionation. Many contaminating non-OMPs were only detected at emPAI cutoff levels below 0.25. For example, using an emPAI threshold of 0.25, in method 2 fractions 73% (65/89) of all OMP emPAIs, but only 46% (192/418) of all non-OMP emPAIs were obtained. This indicates that most non-OMPs were only present in low abundance and are only minor contaminants of the OMFs. This finding is in agreement with SDS-PAGE analysis, which similarly showed that known outer membrane protein bands were prominent (Figure 3). And strikingly, method 2, 3, and 5 OMFs contained significantly less non-OMPs than OMFs prepared with method 1 and 4 (Figure 4B), suggesting that these procedures significantly improved sample quality.

Taken together, all methods yielded a considerable number of identified OMPs, with method 2 being the most specific outer membrane fractionation method as it yields a slightly higher number of identified OMPs than the other methods and results in the lowest amount of contaminating proteins from other cellular compartments. Method 3 is comparably specific but is much more tedious to perform.

Discrepancies between Bioinformatics and Proteomics. We found several discrepancies between numbers of predicted OMPs and experimentally identified ones. Maximally 45% of the PSORTb v.3.0.2 predicted OMPs could be identified (method 4 and 5). One reason for this finding could be that only a small part of predicted OMPs is expressed under the growth conditions we used in our studies. Furthermore, some bacterial membrane proteins are known to have intramolecular amide bonds.[51] Such intramolecular bonds make the identification in mass spectrometric analysis difficult and increase the risk of overlooking those proteins. And finally, one cannot exclude to miss very low abundance OMPs during any one of the steps in OM proteome analysis. Moreover, one has to consider the limitations of the PSORTb v.3.0.2 predictions. A number of identified proteins were annotated as "unknown location" by PSORTb v.3.0.2. Probably, some of those proteins are situated in the OM *in vivo*, and mass spectrometry can help to identify these cases and, in the long run, to improve the predictions.

Subfractionation of All Cellular Compartments using Methods 4 and 5. While method 1, 2, and 3 were established to enrich the OM of Gram-negative bacteria in a fast and uncomplicated manner only, methods 4 and 5 were created to separate all cellular compartments from one origin culture. Figure 5A shows a SDS-PAGE analysis of all cellular subfractions isolated using method 4 and 5. All fractions contained a variety of protein bands with corresponding molecular weights ranging from approximately 10 to 100 kDa, with all subcellular fractions showing remarkable differences in their band pattern.

The quality of the isolated cellular compartments was checked by immunoblotting using different antibodies against marker proteins (Figure 5B). We used anti-GroEL antibodies to identify a cytoplasmic marker protein, anti-TonB antibodies to detect a cytoplasmic membrane marker, anti-Mbp antibodies for the detection of a periplasmic protein marker, and anti-OmpX antibodies to detect an OMP marker in OMFs. The fractions obtained with method 5 showed a higher grade of purity when compared to fractions obtained with method 4. But in most fractions, there were also marker proteins of other cellular compartments detectable, which indicated an incomplete separation of the compartments. Due to the reasons mentioned above and supported by previous OM separation studies with Gram-negative bacteria,[36] we assume that a more effective separation of proteins from different compartments is not possible. The cytoplasmic marker protein GroEL was detectable in all method 4 cellular subfractions and surprisingly

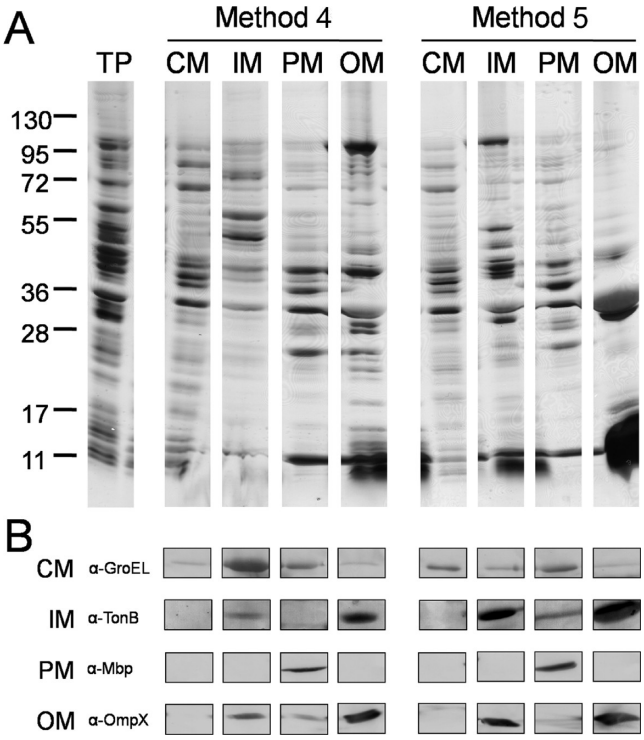


Figure 5. Analyis of subcellular fractions prepared with methods 4 and 5. (A) Approximately 10 μg of *E. coli* BL21 total protein (TP), cytoplasm (CM), cytoplasmic membrane (IM), periplasm (PM) and outer membrane (OM) were separated by 8−18% SDS-PAGE. The positions of molecular mass standards in kDa are shown on the left of the silver-stained gel. (B) Immunoblots of the subcellular fractions using antibodies against marker proteins of the single cell compartments. α-GroEL antibodies was used to detect cytoplasmic (CM) marker protein, α-TonB antibodies to detect cytoplasmic membrane (IM) marker protein, α-Mbp to detect periplasmic (PM) marker proteins, and α-OmpX to detect outer membrane (OM) marker protein.

strongest in the cytoplasmic membrane fraction and the periplasmic fraction. In method 5 subfractions, GroEL was detectable in cytoplasmic and periplasmic fractions. The marker protein for the cytoplasmic membrane, TonB, displayed a stronger signal in method 4 OMFs than in cytoplasmic membrane fractions. In method 5 fractions, TonB was similarly detectable in both cytoplasmic and outer membrane fractions. The periplasmic marker Mbp was detected in none but the periplasmic fractions for both methods 4 and 5. This indicated that periplasmic contaminations in the other fractions were rather low. The OM marker OmpX yielded strongest signals in OMFs for both methods 4 and 5, but there were also slight signals detectable in cytoplasmic membrane fractions and periplasmic fractions derived with both methods.

Obviously, cellular subfractions represent rather an enrichment of proteins of the respective compartment than a sharp separation. This is partly due to the protein properties on which subcellular fractionation is based, such as the solubility in defined detergents, or their integration into or association with membrane vesicles of a certain density. These properties often are blurred between proteins from different compartments. Furthermore, during cell lysis membranes of Gram-negative bacteria appear to aggregate, also based on the fact that many bacterial proteins span or at least are in contact with both membranes. In addition, protein aggregates and large complexes, such as ribosomes and GroEL, can coprecipitate with membrane fractions during ultracentrifugation (e.g., see GroEL protein in inner and outer membrane fractions, Figure 5B).

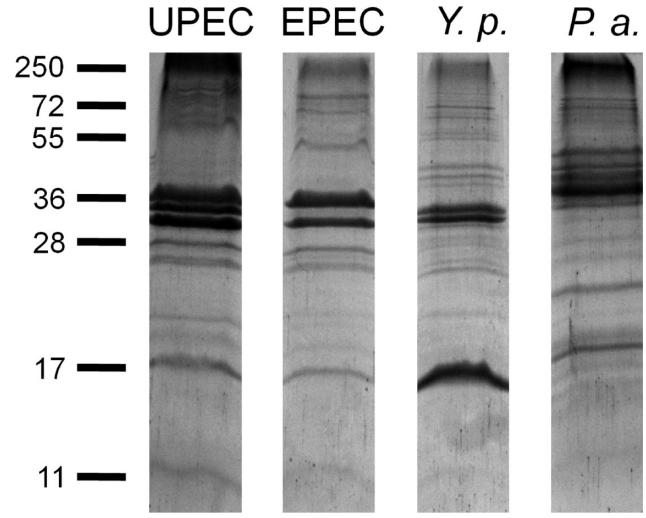


Figure 6. SDS-PAGE analysis of outer membrane fractions of different pathogenic Gram-negative bacteria. Outer membranes of uropathogenic *E. coli* 536 (UPEC), enteropathogenic *E. coli* 2348/69 (EPEC), *Yersinia pseudotuberculosis* (*Y.p.*), and *Pseudomonas aeruginosa* PAO1 (*P.a.*) were separated using method 2. Approximately 10 μg of outer membranes were separated by 15% SDS-PAGE and stained with colloidal Coomassie. The positions of molecular mass standards in kDa are shown in the left.

In spite of the incomplete separation, subfractionation of bacterial cells is often indispensible, for example to evaluate the localization of native or recombinant proteins whose localization prediction is unclear. For such purposes, the subfractionation methods in our study (method 4 and 5) represent feasible techniques. But, for a precise result, the enrichment of a protein in a certain subcellular fraction needs to be taken into account, rather than its mere presence therein.

Outer Membrane Isolation from Pathogens. To obtain information on the applicability of OM fractionation procedures to other Gram-negative bacteria, we selected four pathogenic species from the γ-proteobacteria for our studies. We chose to include the uropathogenic strain *E. coli* 536 (UPEC), the enteropathogenic strain *E. coli* 2348/69 (EPEC), the opportunistic pathogenic strain *Pseudomonas aeruginosa* PA01 (*P.a.*) and *Yersinia pseudotuberculosis* IP32953 (*Y.p.*). We decided to prepare OMFs of those species using method 2, because this method is the most specific one (see above). Furthermore, this method represents a fast and easy protocol that is mostly independent of membrane properties such as lipid composition and vesicle density, which could vary between different species and could lead to further difficulties in OM separation methods, such as selective detergent treatment (as used in methods 1 and 4) and density gradient centrifugation (as used in method 5). An SDS-PAGE analysis of approximately 10 μg of OMFs from the four pathogens is shown in figure 6. The band patterns of the OMs of all four pathogens show clear differences. OMFs of all four species have dominant protein bands at a molecular weight of around 35 kDa. The OMFs of the UPEC and the EPEC strain match in most of the protein bands, which is not a surprise as these strains are closely related. Note, though, that the small differences between these strains probably relate to their different lifestyle.

Outer Membrane Proteomics of Pathogens. In Gram-negative pathogens most pathogenicity factors are located in

Table 3. Outer Membrane Proteome of Pathogens Using Method 2[a]

E. coli 536		*E. coli* 2348/69		*P. aeruginosa* PA01		*Y. pseudotuberculosis* IP32953	
gi number	gene description	gi number	gene description	gi number	gene description	gi number	gene description
gil110641798	major outer membrane lipoprotein precursor	gil215486852	murein lipoprotein	gil15596974	major porin and structural outer membrane porin OprF precursor	gil51595793	outer membrane protein A
gil110641146	outer membrane protein A	gil215486075	outer membrane protein A	gil15598049	Outer membrane lipoprotein OprI precursor	gil51596631	major outer membrane lipoprotein
gil110642425	outer membrane porin protein C	gil215487434	outer membrane porin protein C	gil15596155	basic amino acid, basic peptide and imipenem outer membrane porin OprD precursor	gil51597183	attachment invasion locus protein
gil110642039	outer membrane porin protein LC precursor	gil215485218	organic solvent tolerance protein	gil15598995	hypothetical protein PA3800	gil51596958	long-chain fatty acid outer membrane transporter
gil110640396	outer membrane protein assembly factor YaeT	gil215485389	outer membrane phosphoporin protein E	gil15595488	anaerobically induced outer membrane porin OprE precursor	gil51597310	outer membrane protein assembly factor YaeT
gil110642547	long-chain fatty acid outer membrane transporter	gil215488368	outer membrane channel protein	gil15597956	outer membrane protein precursor	gil51596611	putative lipoprotein
gil110641126	outer membrane protein F	gil215488970	intimin EaeA	gil15595608	twitching motility protein PilJ	gil51594987	organic solvent tolerance protein
gil110643281	outer membrane channel protein	gil215489374	maltoporin	gil15600305	esterase EstA	gil51597698	outer membrane channel protein
gil110640537	putative autotransporter	gil215485338	outer membrane protein assembly factor YaeT	gil15599096	Fe(III) dicitrate transport protein FecA	gil51597154	outer membrane protein assembly complex subunit YfgL
gil110640975	putative pectinesterase	gil215487862	outer membrane protein assembly complex subunit YfgL	gil15596250	hypothetical protein PA1053	gil51595605	outer membrane porin protein C
gil110644375	maltoporin	gil215485787	putative pectinesterase	gil15600167	outer membrane protein precursor	gil51596755	putative lipoprotein
gil110641339	outer membrane porin protein LC precursor	gil215276226	conjugal transfer surface exclusion protein TraT	gil15596485	outer membrane protein precursor	gil51597098	lipoprotein
gil110642677	outer membrane protein assembly complex subunit YfgL	gil215485763	peptidoglycan-associated outer membrane lipoprotein	gil15599870	putative TonB-dependent receptor	gil51595449	LPS-assembly lipoprotein RplB
gil110640949	peptidoglycan-associated outer membrane lipoprotein	gil215488837	putative outer membrane lipoprotein	gil15599262	outer membrane protein OprG precursor	gil51594804	hypothetical protein YPTB0452
gil110643800	putative outer membrane lipoprotein	gil215486054	outer membrane protein F	gil15598888	outer membrane protein precursor	gil51596861	outer membrane protein X
gil110640371	ferrichrome outer membrane transporter	gil215485492	nucleoside channel, receptor of phage T6 and colicin K	gil15599566	metalloproteinase outer membrane protein precursor	gil51597875	outer membrane lipoprotein
gil110642883	lipoprotein NlpD	gil215486101	hypothetical protein E2348C_0969	gil15595238	hypothetical protein PA0040	gil51595504	peptidoglycan-associated outer membrane lipoprotein
gil110644585	putative outer membrane protein	gil215485313	ferrichrome outer membrane transporter	gil15598301	general secretion pathway protein D	gil51595775	porin
gil110640268	organic solvent tolerance protein	gil215489635	N-acetylnuraminic acid outer membrane channel protein	gil15597594	ferripyoverdine receptor	gil51595195	outer membrane protein assembly complex subunit YfiO
gil110641582	outer membrane protein N precursor	gil215487781	lipoprotein	gil15596170	peptidoglycan associated lipoprotein OprL precursor	gil51596382	putative lipoprotein
gil110640477	outer membrane phosphoporin protein E	gil215488062	lipoprotein NlpD	gil15598278	glycine betaine transmethylase		
gil110640672	nucleoside-specific channel-forming protein tsx precursor	gil215487557	predicted lipoprotein	gil15596168	TolA protein		
gil110641763	outer membrane lipoprotein	gil215487934	outer membrane protein assembly complex subunit YfiO	gil15597487	glucose-sensitive porin		
gil110640417	outer membrane lipoprotein	gil215486100	predicted exopolysaccharide export protein	gil15595624	major intrinsic multiple antibiotic resistance efflux outer membrane protein OprM precursor		

Table 3. Continued

E. coli 536		*E. coli* 2348/69		*P. aeruginosa* PA01		*Y. pseudotuberculosis* IP32953	
gi number	gene description	gi number	gene description	gi number	gene description	gi number	gene description
gi|110642758	outer membrane protein assembly complex subunit YfiO	gi|215486818	outer membrane lipoprotein	gi|15598985	putative copper transport outer membrane porin OprC precursor		
gi|110642070	putative outer membrane pore protein	gi|215486640	outer membrane pore protein N, nonspecific	gi|15595360	histidine porin OpdC		
gi|110642651	lipoprotein	gi|215487556	long-chain fatty acid outer membrane transporter	gi|15597718	outer membrane protein precursor CzcC		
gi|229560204	vitamin B12/ cobalamin outer membrane transporter						
gi|110642548	VacJ lipoprotein precursor						
gi|110643163	putative autotransporter						
gi|162138277	murein transglycosylase A						
gi|110643251	putative outer membrane lipoprotein						

[a] This table includes all proteins identified with nanoLC−MS/MS which were annotated as outer membrane proteins by PSORTb v.3.0.2.

the OM, and thus, make OM proteomics indispensible for the identification of therapeutic and diagnostic targets. OM proteomics studies of Gram-negative pathogens, for example, for *Pseudomonas aeruginosa*, have been performed before, using fractionation methods similar to method 2, followed by 2D gel electrophoresis.[16,52] The total number of identified OMPs in published outer membrane proteomics studies of pathogens is often unsatisfactory compared to studies of *E. coli* strains, e.g. in the case of *Yersinia* and *Pseudomonas*.[53−56] To investigate whether the OM proteomics approach using method 2 followed by 1D SDS-PAGE is generally applicable to different Gram-negative pathogens, we analyzed the OMFs of four pathogens in the same manner as described above. Ten μg of OMF was applied to an SDS-PAGE and sliced gel pieces were digested with trypsin. Peptides were analyzed using a nanoLC-ESI ion trap mass spectrometer and proteins were identified using the Mascot search engine as described above. The identified proteins are listed in table S2 (supplemental data). We identified 133 proteins in the OMFs of *E. coli* 536, 83 proteins in the OMFs of *E. coli* 2348/69, 101 proteins in the OMFs of *P. aeruginosa*, and 96 proteins in the OMFs of *Y. pseudotuberculosis*. The subcellular localization of these proteins was predicted by PSORTb v.3.0.2 We identified 32 OMPs from *E. coli* 536, 27 OMPs from *E. coli* 2348/69, 27 OMPs from *P. aeruginosa* and 20 OMPs from *Y. pseudotuberculosis* (Table 3). Interestingly, the numbers of identified OMPs from the uropathogenic *E. coli* strain 536 and of *P. aeruginosa* PA01 were slightly higher than the number of OMPs identified from *E. coli* BL21. As many pathogenicity factors are surface-localized, it is tempting to assume that these additional proteins are interesting targets for future experiments.

Conclusions

In this study we investigated the practicability of 5 different subfractionation methods for bacterial cells. Subfractionation methods can be varied according to the requirements of the experiment, ranging from a fast and easy OM preparation (method 1, 2 and 3) to an extensive cellular subfractionation of Gram-negative bacteria (method 4 and 5), with methods 2 and 5 showing best results in OM preparation and cellular subfractionation, respectively.

We demonstrated that the OM samples obtained with these methods can be used in mass spectrometry experiments after simple, one-dimensional SDS-PAGE separation. The methods allow the identification of a high number of OMPs, although contaminations by proteins of other cell compartments in the preparations are inevitable. The quality of the data (i.e., the amount of OMPs identified) after tryptic digestion of gel segments is comparable to results obtained by "classical" two-dimensional separation.

Fractionation method 2 yielded the highest amount of identified OMPs, and yielded significantly fewer contaminating proteins than OMFs obtained by the other methods. Method 2 includes a washing step with chaotropic reagents, which seems to significantly increase the purity of the OM fractions. Effective OM enrichment is indispensible for proteomic studies e.g. under varying growth conditions, and in principle allows the comparison of expression levels of certain proteins.

Finally, we show the applicability of method 2 to different Gram-negative, pathogenic species, which is an important prerequisite for extensive proteomics studies aiming at biomarker discovery, and at the development of antimicrobial drugs and vaccines.[3]

Acknowledgment. We thank Andrei Lupas for continuing support, Philipp Oberhettinger (Medizinische Mikrobiologie, Tübingen) for help during the work with pathogens, and Heike Hoffmann for technical support. Furthermore, the authors thank Ulrich Dobrindt (Institut für Molekulare Infektionsbiologie, Würzburg) for kindly providing us with the pathogenic *E. coli* strains 536 and 2348/69, and the Medizinische Mikrobiologie, Tübingen, headed by Prof. Ingo Autenrieth, for providing us with the pathogenic strains *Yersinia pseudotuberculosis* IP32953 and *Pseudomonas aeruginosa* PAO1. This work was supported by the Bill and Melinda Gates Foundation - Grand Challenges and Explorations (grant ID 51949).

Supporting Information Available: Supplementary Tables S1 and S2. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Vitek, O. Getting started in computational mass spectrometry-based proteomics. *PLoS Comput. Biol.* **2009**, *5* (5), e1000366.
(2) Ovsyannikova, I. G.; Johnson, K. L.; Bergen, H. R., 3rd; Poland, G. A. Mass spectrometry and peptide-based vaccine development. *Clin. Pharmacol. Ther.* **2007**, *82* (6), 644–52.
(3) Sivick, K. E.; Mobley, H. L. An "omics" approach to uropathogenic *Escherichia coli* vaccinology. *Trends Microbiol* 2009.
(4) Jungblut, P. R. Proteome analysis of bacterial pathogens. *Microbes Infect.* **2001**, *3* (10), 831–40.
(5) Molloy, M. P. Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Anal. Biochem.* **2000**, *280* (1), 1–10.
(6) Rabilloud, T. Membrane proteins and proteomics: love is possible, but so difficult. *Electrophoresis* **2009**, *30* (1), S174–80.
(7) Santoni, V.; Molloy, M.; Rabilloud, T. Membrane proteins and proteomics: un amour impossible. *Electrophoresis* **2000**, *21* (6), 1054–70.
(8) Neu, H. C.; Heppel, L. A. The release of enzymes from *Escherichia coli* by osmotic shock and during the formation of spheroplasts. *J. Biol. Chem.* **1965**, *240* (9), 3685–92.
(9) Köster, W.; Braun, V. Iron-hydroxamate transport into *Escherichia coli* K12: localization of FhuD in the periplasm and of FhuB in the cytoplasmic membrane. *Mol. Gen. Genet.* **1989**, *217* (2−3), 233–9.
(10) Owen, P.; Kaback, H. R. Antigenic architecture of membrane vesicles from *Escherichia coli*. *Biochemistry* **1979**, *18* (8), 1422–6.
(11) Dickie, P.; Weiner, J. H. Purification and characterization of membrane-bound fumarate reductase from anaerobically grown *Escherichia coli*. *Can. J. Biochem.* **1979**, *57* (6), 813–21.
(12) Bertero, M. G.; Rothery, R. A.; Palak, M.; Hou, C.; Lim, D.; Blasco, F.; Weiner, J. H.; Strynadka, N. C.; Insights into the respiratory electron transfer pathway from the structure of nitrate reductase, A. *Nat. Struct. Biol.* **2003**, *10* (9), 681–7.
(13) Osborn, M. J.; Gander, J. E.; Parisi, E.; Carson, J. Mechinism of assembly of the outer membrane of *Salmonella typhimurium*. Isolation and characterization of cytoplasmic and outer membrane. *J. Biol. Chem.* **1972**, *247* (12), 3962–72.
(14) Schnaitman, C. A. Effect of ethylenediaminetetraacetic acid, Triton X-100, and lysozyme on the morphology and chemical composition of isolate cell walls of *Escherichia coli*. *J. Bacteriol.* **1971**, *108* (1), 553–63.
(15) Fujiki, Y.; Hubbard, A. L.; Fowler, S.; Lazarow, P. B. Isolation of intracellular membranes by means of sodium carbonate treatment: application to endoplasmic reticulum. *J. Cell Biol.* **1982**, *93* (1), 97–102.
(16) Molloy, M. P.; Herbert, B. R.; Slade, M. B.; Rabilloud, T.; Nouwens, A. S.; Williams, K. L.; Gooley, A. A. Proteomic analysis of the *Escherichia coli* outer membrane. *Eur. J. Biochem.* **2000**, *267* (10), 2871–81.
(17) Weiner, J. H.; Li, L. Proteome of the *Escherichia coli* envelope and technological challenges in membrane proteome analysis. *Biochim. Biophys. Acta* **2008**, *1778* (9), 1698–713.
(18) Yu, N. Y.; Wagner, J. R.; Laird, M. R.; Melli, G.; Rey, S.; Lo, R.; Dao, P.; Sahinalp, S. C.; Ester, M.; Foster, L. J.; Brinkman, F. S. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **2010**, *26* (13), 1608–15.
(19) Remmert, M.; Linke, D.; Lupas, A. N.; Soding, J. HHomp--prediction and classification of outer membrane proteins. *Nucleic Acids Res.* **2009**, *37* (Web Server issue), W446–51.
(20) Berven, F. S.; Flikka, K.; Jensen, H. B.; Eidhammer, I. BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res.* **2004**, *32*, W394–9.
(21) Straley, S. C.; Brubaker, R. R. Cytoplasmic and membrane proteins of yersiniae cultivated under conditions simulating mammalian intracellular environment. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78* (2), 1224–8.
(22) Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **1970**, *227* (5259), 680–5.
(23) Blum, H.; Beier, H.; Gross, H. J. Improved silver staining of plant-proteins, RNA and DNA in polyacrylamide gels. *Electrophoresis* **1987**, *8* (2), 93–99.
(24) Winkler, C.; Denker, K.; Wortelkamp, S.; Sickmann, A. Silver- and Coomassie-staining protocols: detection limits and compatibility with ESI MS. *Electrophoresis* **2007**, *28* (12), 2095–9.
(25) Mehta, P. Semi-dry protein transfer and immunodetection of P-selectin using an antibody to its C-terminal tag. *Methods Mol. Biol.* **2009**, *536*, 229–35.
(26) Towbin, H.; Staehelin, T.; Gordon, J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U.S.A.* **1979**, *76* (9), 4350–4.
(27) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **1996**, *68* (5), 850–8.
(28) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2* (8), 1896–906.
(29) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–402.
(30) Pautsch, A.; Schulz, G. E. High-resolution structure of the OmpA membrane domain. *J. Mol. Biol.* **2000**, *298* (2), 273–82.
(31) Buchanan, S. K.; Smith, B. S.; Venkatramani, L.; Xia, D.; Esser, L.; Palnitkar, M.; Chakraborty, R.; van der Helm, D.; Deisenhofer, J. Crystal structure of the outer membrane active transporter FepA from *Escherichia coli*. *Nat. Struct. Biol.* **1999**, *6* (1), 56–63.
(32) Galdiero, S.; Galdiero, M.; Pedone, C. beta-Barrel membrane bacterial proteins: structure, function, assembly and interaction with lipids. *Curr. Protein Pept. Sci.* **2007**, *8* (1), 63–82.
(33) Schulz, G. E. The structure of bacterial outer membrane proteins. *Biochim. Biophys. Acta* **2002**, *1565* (2), 308–17.
(34) Braun, V. Covalent lipoprotein from the outer membrane of *Escherichia coli*. *Biochim. Biophys. Acta* **1975**, *415* (3), 335–77.
(35) Hayashi, S.; Wu, H. Lipoproteins in bacteria. *J. Bioenerg. Biomembr.* **1990**, *22* (3), 451–71.
(36) Hobb, R. I.; Fields, J. A.; Burns, C. M.; Thompson, S. A. Evaluation of procedures for outer membrane isolation from *Campylobacter jejuni*. *Microbiology* **2009**, *155* (Pt 3), 979–88.
(37) Grobner, S.; Linke, D.; Schutz, W.; Fladerer, C.; Madlung, J.; Autenrieth, I. B.; Witte, W.; Pfeifer, Y. Emergence of carbapenem-non-susceptible extended-spectrum beta-lactamase-producing *Klebsiella pneumoniae* isolates at the university hospital of Tübingen, Germany. *J. Med. Microbiol.* **2009**, *58* (Pt 7), 912–22.
(38) Lee, E. Y.; Bang, J. Y.; Park, G. W.; Choi, D. S.; Kang, J. S.; Kim, H. J.; Park, K. S.; Lee, J. O.; Kim, Y. K.; Kwon, K. H.; Kim, K. P.; Gho, Y. S. Global proteomic profiling of native outer membrane vesicles derived from *Escherichia coli*. *Proteomics* **2007**, *7* (17), 3143–53.
(39) Fountoulakis, M.; Gasser, R. Proteomic analysis of the cell envelope fraction of *Escherichia coli*. *Amino Acids* **2003**, *24* (1−2), 19–41.
(40) Hartinger, J.; Stenius, K.; Hogemann, D.; Jahn, R. 16-BAC/SDS-PAGE: a two-dimensional gel electrophoresis system suitable for the separation of integral membrane proteins. *Anal. Biochem.* **1996**, *240* (1), 126–33.
(41) Chevallet, M.; Santoni, V.; Poinas, A.; Rouquie, D.; Fuchs, A.; Kieffer, S.; Rossignol, M.; Lunardi, J.; Garin, J.; Rabilloud, T. New zwitterionic detergents improve the analysis of membrane proteins by two-dimensional electrophoresis. *Electrophoresis* **1998**, *19* (11), 1901–9.
(42) Molloy, M. P.; Herbert, B. R.; Walsh, B. J.; Tyler, M. I.; Traini, M.; Sanchez, J. C.; Hochstrasser, D. F.; Williams, K. L.; Gooley, A. A. Extraction of membrane proteins by differential solubilization for separation using two-dimensional gel electrophoresis. *Electrophoresis* **1998**, *19* (5), 837–44.
(43) Fountoulakis, M.; Takacs, B. Effect of strong detergents and chaotropes on the detection of proteins in two-dimensional gels. *Electrophoresis* **2001**, *22* (9), 1593–602.
(44) Cordwell, S. J.; Thingholm, T. E. Technologies for plasma membrane proteomics. *Proteomics* 2009.
(45) Herbert, B. Advances in protein solubilisation for two-dimensional electrophoresis. *Electrophoresis* **1999**, *20* (4−5), 660–3.
(46) McDonough, J.; Marban, E. Optimization of IPG strip equilibration for the basic membrane protein mABC1. *Proteomics* **2005**, *5* (11), 2892–5.
(47) Lopez-Campistrous, A.; Semchuk, P.; Burke, L.; Palmer-Stone, T.; Brokx, S. J.; Broderick, G.; Bottorff, D.; Bolch, S.; Weiner, J. H.; Ellison, M. J. Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth. *Mol. Cell. Proteomics* **2005**, *4* (8), 1205–9.

(48) Walters, M. S.; Mobley, H. L. Identification of uropathogenic *Escherichia coli* surface proteins by shotgun proteomics. *J. Microbiol. Methods* 2009.
(49) Cirulli, C.; Marino, G.; Amoresano, A. Membrane proteome in Escherichia coli probed by MS3 mass spectrometry: a preliminary report. *Rapid Commun. Mass Spectrom.* **2007**, *21* (14), 2389–97.
(50) Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **2005**, *4* (9), 1265–72.
(51) Budzik, J. M.; Poor, C. B.; Faull, K. F.; Whitelegge, J. P.; He, C.; Schneewind, O. Intramolecular amide bonds stabilize pili on the surface of bacilli. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (47), 19992–7.
(52) Nouwens, A. S.; Cordwell, S. J.; Larsen, M. R.; Molloy, M. P.; Gillings, M.; Willcox, M. D.; Walsh, B. J. Complementing genomics with proteomics: the membrane subproteome of Pseudomonas aeruginosa PAO1. *Electrophoresis* **2000**, *21* (17), 3797–809.
(53) Nouwens, A. S.; Willcox, M. D.; Walsh, B. J.; Cordwell, S. J. Proteomic comparison of membrane and extracellular proteins from invasive (PAO1) and cytotoxic (6206) strains of *Pseudomonas aeruginosa*. *Proteomics* **2002**, *2* (9), 1325–46.
(54) Pieper, R.; Huang, S. T.; Robinson, J. M.; Clark, D. J.; Alami, H.; Parmar, P. P.; Perry, R. D.; Fleischmann, R. D.; Peterson, S. N. Temperature and growth phase influence the outer-membrane proteome and the expression of a type VI secretion system in *Yersinia pestis*. *Microbiology* **2009**, *155* (Pt 2), 498–512.
(55) Peng, X.; Xu, C.; Ren, H.; Lin, X.; Wu, L.; Wang, S. Proteomic analysis of the sarcosine-insoluble outer membrane fraction of *Pseudomonas aeruginosa* responding to ampicilin, kanamycin, and tetracycline resistance. *J. Proteome Res.* **2005**, *4* (6), 2257–65.
(56) Blonder, J.; Goshe, M. B.; Xiao, W.; Camp, D. G., 2nd; Wingerd, M.; Davis, R. W.; Smith, R. D. Global analysis of the membrane subproteome of *Pseudomonas aeruginosa* using liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **2004**, *3* (3), 434–44.

PR1002438

frontiers in
MICROBIOLOGY

METHODS ARTICLE
published: 08 November 2011
doi: 10.3389/fmicb.2011.00218

ClubSub-P: cluster-based subcellular localization prediction for Gram-negative bacteria and archaea

Nagarajan Paramasivam and Dirk Linke*

Department I Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany

Edited by:
Martin G. Klotz, University of North Carolina at Charlotte, USA

Reviewed by:
Loren Hauser, Oak Ridge National Laboratory, USA
Uli Stingl, King Abdullah University of Science and Technology, Saudi Arabia

***Correspondence:**
Dirk Linke, Department I Protein Evolution, Max Planck Institute for Developmental Biology, Spemannstr. 35, D-72076 Tübingen, Germany.
e-mail: dirk.linke@tuebingen.mpg.de

The subcellular localization (SCL) of proteins provides important clues to their function in a cell. In our efforts to predict useful vaccine targets against Gram-negative bacteria, we noticed that misannotated start codons frequently lead to wrongly assigned SCLs. This and other problems in SCL prediction, such as the relatively high false-positive and false-negative rates of some tools, can be avoided by applying multiple prediction tools to groups of homologous proteins. Here we present ClubSub-P, an online database that combines existing SCL prediction tools into a consensus pipeline from more than 600 proteomes of fully sequenced microorganisms. On top of the consensus prediction at the level of single sequences, the tool uses clusters of homologous proteins from Gram-negative bacteria and from Archaea to eliminate false-positive and false-negative predictions. ClubSub-P can assign the SCL of proteins from Gram-negative bacteria and Archaea with high precision. The database is searchable, and can easily be expanded using either new bacterial genomes or new prediction tools as they become available. This will further improve the performance of the SCL prediction, as well as the detection of misannotated start codons and other annotation errors. ClubSub-P is available online at http://toolkit.tuebingen.mpg.de/clubsubp/

Keywords: subcellular localization prediction, signal peptide, clustering, protein homology, start codon prediction

INTRODUCTION

Gram-negative bacteria have a multi-layered cell envelope, which consists of a symmetrical phospholipid bilayer (the cytoplasmic or inner membrane, IM) and an asymmetrical bilayer comprised of phospholipids and lipopolysaccharides (the outer membrane, OM). These membranes are separated by the periplasmic space, which contains a thin peptidoglycan layer as a cell wall (Gardy and Brinkman, 2006; Bos et al., 2007). The IM is the boundary for the cytosol; thus the Gram-negative cell consists of four compartments (cytosol, IM, periplasm, OM). Each subcellular compartment contains a defined set of proteins to fulfill distinct tasks.

To perform their functions at their native subcellular localization (SCL), newly synthesized proteins must be sorted and transported to their respective subcellular compartments. While most of the newly synthesized proteins remain in the cytoplasm, other proteins are inserted into the cytoplasmic membrane via the signal recognition particle (SRP) and YidC pathways. Proteins are targeted to the cytoplasmic membrane via the SRP pathway. YidC acts like an additional insertase to fold and assemble a defined subset of these proteins in the cytoplasmic membrane (Luirink et al., 2005). Proteins with native functions in the periplasmic space and in the OM are secreted across the cytoplasmic membrane into the periplasmic space by the Sec, TAT, or Holin (which secretes autolytic enzymes during cell death; Saier et al., 2008) secretory pathways. From the periplasm some proteins are further translocated to the OM or across the OM via Type II secretion systems (T2SS), T5SS, T7SS, and T8SS. Secretion systems such as T1SS, T3SS, T4SS, and T6SS span both membranes and can secrete proteins from the cytoplasm directly into the extracellular space or even into the host cytoplasm (Desvaux et al., 2009).

The general secretion system (Sec; Desvaux et al., 2009) is the most common pathway; it is conserved in all living organisms. In Gram-negative bacteria, it translocates unfolded proteins across the cytoplasmic membrane into the periplasmic space. The Sec translocon recognizes signal sequences present at the N-terminus of its substrate proteins. These general Sec signals are highly conserved and consist of a positively charged N-terminal region (n-region), a hydrophobic central region (h-region), and a polar C-terminal region (c-region; Nielsen et al., 1997). Alternatively, some folded proteins use the twin-arginine translocation (TAT) pathway for secretion across the cytoplasmic membrane, which recognizes its substrates through a modified general signal peptide with an additional RRXFL motif found between the n-region and h-region (Bendtsen et al., 2005). Typically, TAT signal peptides are longer than general signal peptides. The secretion of lipoproteins is accomplished by another modification of the general Sec signal peptide pathway. Here a cysteine residue follows immediately after the signal peptide cleavage site; this signal peptide is recognized and cleaved by lipoprotein signal peptidase (SPaseII or Lsp) after the N-terminal cysteine is modified with a lipid moiety, which anchors the protein to the membrane. Finally, an additional fatty acid is attached to the new N-terminus (Juncker et al., 2003). These proteins are then either retained at the cytoplasmic membrane or translocated to the OM by the Lol lipoprotein-sorting pathway (Lewenza et al., 2008). Although this sorting is assumed to be based on the residue at the +2 position after the cleavage site (Seydel et al., 1999), it has been shown that residues at +3 and

+4 also play important roles in the sorting of these proteins in *Pseudomonas aeruginosa* (Lewenza et al., 2008). So far, the detailed patterns of lipoprotein-sorting remain unclear. A number of specialized secretion systems exist, each one typically translocating only a small subset of proteins.

The SCL of proteins provides important clues to their function in the cell. Determining the SCL of proteins by experimental means is accurate but time-consuming and expensive. As a result of new and more efficient sequencing technologies, the number of newly deposited sequences is increasing exponentially, while the number of proteins annotated with experimentally verified SCL stagnates. Thus, computational SCL prediction is important and has become indispensible in protein research, e.g., for genome-wide SCL studies. There are two types of SCL prediction tools. One type is predicting only the features specific to localizations, such as signal peptides (Nielsen et al., 1997; Rose et al., 2002; Juncker et al., 2003; Bendtsen et al., 2004, 2005; Hiller et al., 2004; Käll et al., 2004; Bos et al., 2007; Szabó et al., 2007; Arnold et al., 2009; Bagos et al., 2009; Löwer and Schneider, 2009), transmembrane helices (TMHs; Krogh et al., 2001; Tusnady and Simon, 2001; Käll et al., 2004), or transmembrane β-barrels (TMBBs; Berven et al., 2004; Remmert et al., 2009). The other type is predicting the exact localization of a protein by combining various localization-specific features (Su et al., 2007; Yu et al., 2010) or general features like amino acid composition (Yu et al., 2006), evolutionary information (Rashid et al., 2007), structure conservation information (Su et al., 2007), and gene ontology (Chou and Shen, 2006b).

It has been shown that the combination of different SCL prediction tools increases the quality of the overall prediction significantly (Shen and Burger, 2007; Horler et al., 2009; Giombini et al., 2010; Goudenège et al., 2010). Moreover, Imai and Nakai (2010) recently reported that homology-based methods perform better even on datasets with a low overall sequence identity cutoff, when compared to state-of-the-art single-sequence SCL predictors. Mah et al. (2010) used clustering information to optimize OM β-barrel protein predictions in seven proteomes of Mycobacteria.

Our interest is predominantly in surface-localized proteins of Gram-negative bacteria that could be exploited for vaccine development. We found most single SCL prediction methods to be either not useful or not sensitive enough for our bioinformatics pipeline. Moreover, we found many proteins with misannotated start codons. These are easily identified from the multiple sequence alignments of homologous proteins but are hard to find on the level of individual sequences. The differences in start codon predictions between orthologous sequences from closely related organisms are typically a result of using different automated gene prediction methods while annotating the sequenced genome (Overbeek et al., 2007). These misannotations are a common source of error in SCL prediction, especially since feature prediction tools based on N-terminal signal peptides depend essentially on accurate annotations of the translation start. Conversely, the TMBB prediction tool BOMP uses a C-terminal β-barrel motif for its predictions and thus relies on correctly sequenced stop codons (Berven et al., 2004).

In this work, we developed a method called cluster-based SCL prediction, or ClubSub-P, which combines different localization-specific features and SCL prediction tools, using rules based on the biology of protein sorting to annotate the SCL for Gram-negative bacterial proteins. In contrast to other general SCL prediction tools, it uses homology information taken from clusters of orthologous proteins from different species to further increase the confidence of the prediction. Since we use information from the whole cluster to increase the confidence, we overcome the problem of misannotation of start codons and thus increase the specificity of the method further. Performance measurements with ClubSub-P show that the additional use of homology information from simple clustering increases the precision of our tool over other state-of-the-art SCL prediction tools. Our tool relies on an expandable database. The constantly increasing number of sequenced genomes will, over time, allow us to cluster more sequences, which will further increase the quality of homology detection and thus, the precision of our predictions. To show how easily the tool can be expanded to whole new organism groups, we have included an additional module for the SCL prediction of archaeal proteins.

MATERIALS AND METHODS

DATASETS

To create the ClubSub-P database (see Database, below), 607 Gram-negative bacterial proteomes (2,331,935 sequences) were downloaded from the NCBI RefSeq genome database[1] in July 2011. A non-redundant dataset was created using CD-HIT (Li and Godzik, 2006) from the above sequences at 40% local sequence identity, and at 80% sequence alignment coverage to the longest sequence in the cluster. The "accurate and slow" mode was used to ensure clustering of proteins into the most similar cluster, which is not given when using the fast mode. Shorter sequences (<40 amino acids) were removed from the dataset for two reasons. First, such short proteins are only annotated in very few bacterial genomes and frequently do not show significant homology to proteins with experimentally verified SCL (Warren et al., 2010). Second, even when there is available experimental data, small proteins are frequently considered fragments and are removed from datasets of many SCL prediction tools (Chou and Shen, 2006a), making a consensus prediction impossible. The final dataset, which we named DB_ClubSub-P, contained 1,911,760 proteins. The list of the downloaded proteomes and the accession numbers of the replicons are given in Data Sheet S1 in Supplementary Material.

We used the Gram-negative bacterial protein sequences from the training dataset of PSORTb v3.0.2[2] (Yu et al., 2010) to test the clustering parameters. This dataset contains 8,227 protein sequences with experimentally determined SCLs and we named it DB_ePSORT.

To obtain a test set for the evaluation of the performance of ClubSub-P, Gram-negative bacterial protein sequences with experimentally verified SCL annotation were extracted from UniProt Release 2011_07 (UniProt-Consortium, 2010). We wrote a parser to extract Gram-negative bacterial protein sequences with literature reference to their SCL annotations, but ignoring

[1] ftp://ftp.ncbi.nih.gov/genomes/Bacteria/
[2] http://www.psort.org/dataset/datasetv3.html

sequences with "potential," "by similarity," or "probable" annotations, sequences labeled as "Fragment," or sequences with "chromatophore" localization.

Sequences with ≤40 aa length were removed from this dataset; we also removed sequences which have more than 40% sequence identity to the PSORTb v.3 training dataset to allow an objective comparison between the tools. Likewise, since the SCL tools used in performance measure do not separately annotate lipoproteins, we removed sequences with "lipid anchor" SCL annotation which leaves 171 sequences for our DB_Experimental dataset.

SUBCELLULAR LOCALIZATION PREDICTION

Subcellular localization prediction using the DB_ClubSub-P dataset was done on two levels. First, we combined different prediction tools as listed in **Table 1** for localization-specific features, and SCL prediction tools based on known biological rules as shown in **Table 2**, to annotate the SCL of each single protein in the DB_ClubSub-P dataset. **Figure 1** displays the procedure in form of flow chart. Second, we clustered all protein sequences and combined their SCL annotations into a consensus SCL prediction for each protein cluster.

Consensus subcellular localization at the protein level

Consensus signal peptide prediction. Signal peptide predictions for Lipoprotein signals, TAT pathway signal peptides, general secretory signal peptides, T3SS signal peptides and T4SS signal peptides were done on all proteins in the DB_ClubSub-P dataset. A lipoprotein prediction was considered positive when

Table 1 | List of SCL and feature specific tools used in the prediction pipeline.

Tools	Features of SCL**	Used for‡	Signal peptide prediction modes	Prediction threshold (default threshold from the predictors)	References
LipoP1.0	SPII	Archaea and Gram−	Gram-negative bacteria	Best prediction: SpII	Juncker et al. (2003)
Tatp 1.0	TAT	Archaea and Gram−	Bacteria	Twin-arginine motif and MaxDscore >0.36	Bendtsen et al. (2005)
TaTFind 1.4	TAT	Archaea and Gram−	Prokaryote	Rules 3a, 3b, or 4*	Rose et al. (2002)
SignalP 3.0-NN	GSP	Archaea and Gram−	Gram-positive and Gram-negative bacteria	MaxDscore >0.44	Bendtsen et al. (2004)
SignalP 3.0-HMM	GSP	Archaea and Gram−	Gram-positive and Gram-negative bacteria	SP probability >0.5	Bendtsen et al. (2004)
Predisi	GSP	Gram−	Gram-negative bacteria	Prediction score >0.5	Hiller et al. (2004)
RPSP	GSP	Gram−	Prokaryote	Positive SP prediction	Plewczynskia et al. (2007)
Phobius	GSP, IMP	Archaea and Gram−	–	Positive SP prediction and TMH prediction	Käll et al. (2004)
TMHMM 2.0.0	IMP	Archaea and Gram−	–	Positive TMH prediction	Krogh et al. (2001)
HMMTOP 2.0	IMP	Archaea and Gram−	–	Positive TMH prediction	Tusnady and Simon (2001)
EffectiveT3	T3SS	Gram−	Gram-negative bacteria	Prediction score ≥0.8$	Arnold et al. (2009)
T3SS_prediction	T3SS	Gram−	Gram-negative bacteria	Prediction score ≥0.8$	Löwer and Schneider (2009)
PSORTb v3.0.2	OMP, LPP, EXT, CW	Archaea and Gram−	–	Final prediction – outer membrane or extracellular or cell wall***	Yu et al. (2010)
CELLO v.2.5	OMP, LPP	Gram−	–	Final prediction – outer membrane***	Yu et al. (2006)
BOMP	OMBB	Gram−	–	Positive prediction (category 1–5)	Berven et al. (2004)
HHomp	OMBB	Gram−	–	OMP probability ≥90$	Remmert et al. (2009)
PRED-SIGNAL	GSP	Archaea	Archaea	Positive "signal" prediction	Bagos et al. (2009)
FlaFind	Prepilin SP	Archaea	Archaea	Positive prepilin signal detection	Szabó et al. (2007)
PilFind	Type IV pilin SP	Gram−	–	Positive pilin signal peptide	Imam et al. (submitted)

**Twin-arginine motif followed by a single charged residue (Rule: 3a, 3b) or basic residue following the twin-arginine and hydrophobic stretch (Rule 4).*

***SPII, lipoprotein signal peptide; TAT, TAT signal peptide; GSP, general signal peptide; CMP, cytoplasmic membrane protein; T3SS, type 3 secretory signal peptide; OMP, outer membrane protein; EXT, extracellular protein; LPP, leaderless periplasmic protein; OMBB, outer membrane β-barrel; Prepilin SP, prepilin signal peptide.*

‡Gram −, Gram-negative bacteria.

****Periplasmic prediction used only when there is no consensus signal peptide prediction.*

$User defined the cutoffs.

Table 2 | Logic for SCL prediction at the protein level.

Features	Lipoprotein SP	Consensus TAT SP	Consensus general SP	Consensus TMH	Consensus TMBB	Consensus T3SS SP or T4SS SP or extracellular
LOCALIZATION						
Cytoplasm	No	No	No	No	No	No
Cytoplasmic membrane	No	No	No	1 or more	No	No
Periplasm	No	Any one of the SP		No	No	No
Lipoprotein	Yes	No	No	No	No	No
Outer membrane	Any one of the SP			No	Yes	No
Extracellular	No	Yes or no		0 or more	No	Yes

the best prediction of LipoP 1.0 (Juncker et al., 2003) was for a signal peptidase II cleavage site. For TAT pathway signal peptide prediction in ClubSub-P, both TatP 1.0 (Bendtsen et al., 2005) and the rule-based predictor TatFind 1.4 (Rose et al., 2002) had to be positive; the cutoff for a positive TatP 1.0 prediction was a MaxD score above 0.36, while TatFind 1.4 requires the presence of the twin-arginine motif and additional sequence features.

Five tools were combined for the consensus prediction of general signal peptides: SignalP-HMM (with a default cutoff of $p = 0.5$), SignalP-NN (with MaxD value above 0.44), Predisi (with a default cutoff of $p = 0.5$), RPSP (with positive signal peptide), or Phobius (with positive signal peptide prediction; Bendtsen et al., 2004; Hiller et al., 2004; Käll et al., 2004). For a positive prediction, three out of five tools were required to be positive; in this case, a consensus SP cleavage site was predicted from the individual cleavage site predictions. Here, Phobius was also used to differentiate between the SP and TMH predictions (see below). If only two tools predict the presence of a signal peptide with zero or one consensus TMHs, the protein's SCL is annotated as "Unknown" to avoid false-positive predictions.

To reduce the false-positive prediction rate of type III signal peptide prediction, positive predictions from both EffectiveT3 (Arnold et al., 2009) and T3SS_prediction (Löwer and Schneider, 2009) were required; predictions with scores $\geq$0.8 were considered as positive type III signal peptides (Burstein et al., 2009). We used a new, unpublished tool named PilFind (Imam et al., submitted) to predict type IV secretion system (T4SS) signals.

If one or more SP were predicted for a protein, it was classified based on the hierarchy described above (**Figure 1**), since there are cases where Lipoprotein or TAT SPs are also predicted as general SPs by general SP prediction tools, and taking into consideration that the accuracy of T3SS and T4SS SP prediction tools is still insufficient.

Consensus transmembrane helix prediction. TMHMM (Krogh et al., 2001), HMMTOP (Tusnady and Simon, 2001), and Phobius (Käll et al., 2004) were used for the prediction of TMHs. For the consensus TMH prediction, we ruled that a helix must be predicted independently by at least 2 of the tools used, over a length of at least 10 residues. Consensus TMH prediction was avoided over the length of previously predicted cleavable signal peptides, because signal peptides are known to be frequently misinterpreted as TMHs by TM prediction tools. The consensus TMH prediction is displayed in **Figure 2**.

Consensus transmembrane β-barrel prediction. We used BOMP (Berven et al., 2004), CELLO (Yu et al., 2006), PSORTb (Yu et al., 2010), and HHomp (Remmert et al., 2009) to predict outer membrane proteins (OMPs). Since classifier-based predictions are faster than sensitive search methods such as HHomp, only BOMP, CELLO, and PSORTb were ran on all the sequences. If any one of BOMP, PSORTb, or CELLO had a positive prediction for OMPs in a cluster (see Subcellular Localization on the Level of Sequence Clusters for details on clustering), we selected a random sequence from the cluster and ran HHomp. When the sequence was predicted as OMP with probability above 90%, we annotated all the sequences in the cluster as OM-localized TMBBs.

Consensus subcellular localization prediction. For the consensus SCL prediction we applied rules based on the biology of protein sorting along with the previously predicted protein features as mentioned in the **Table 2**. The lipoprotein-sorting signal is based on the amino acids after the SPII cleavage site and species-specific (Juncker et al., 2003). Currently there is not sufficient experimental data to postulate a common sorting pattern for all species. Thus, we annotated proteins with lipoprotein signal peptides and without TMHs as "IM/OM lipoprotein." Also, as there is insufficient experimental data available to annotate the extracellular presence of lipoproteins, we didn't analyze the further destination of lipoproteins (Pugsley et al., 1990). Proteins featuring general Sec or TAT signal peptides and without TMHs and TMBBs were annotated as "periplasmic." Proteins predicted to be periplasmic by PSORTb v3.0.2 (Yu et al., 2010) and CELLO v.2.5 (Yu et al., 2006) but without any signal peptide, TMHs and TMBBs were also predicted as periplasmic. Additionally, they were tagged with a note stating that they could be secreted via signal peptide-independent pathways (leaderless pathways). Proteins with one or more consensus TMHs were annotated as "cytoplasmic membrane." Proteins with consensus TMBB prediction containing one of the previously predicted cleaved general, TAT, or lipoprotein signal peptides were annotated as "outer membrane protein," as OMPs are typically secreted by SP-dependent pathways. The SCL of proteins with positive TMBB predictions, but without any signal peptide predictions were annotated as "Unknown". Proteins predicted to be extracellular by PSORTb or predicted to have a T3SS or T4SS signal peptide were annotated as "extracellular." Proteins without TMHs, TMBBs, signal peptide, or extracellular prediction were annotated as "cytoplasmic."

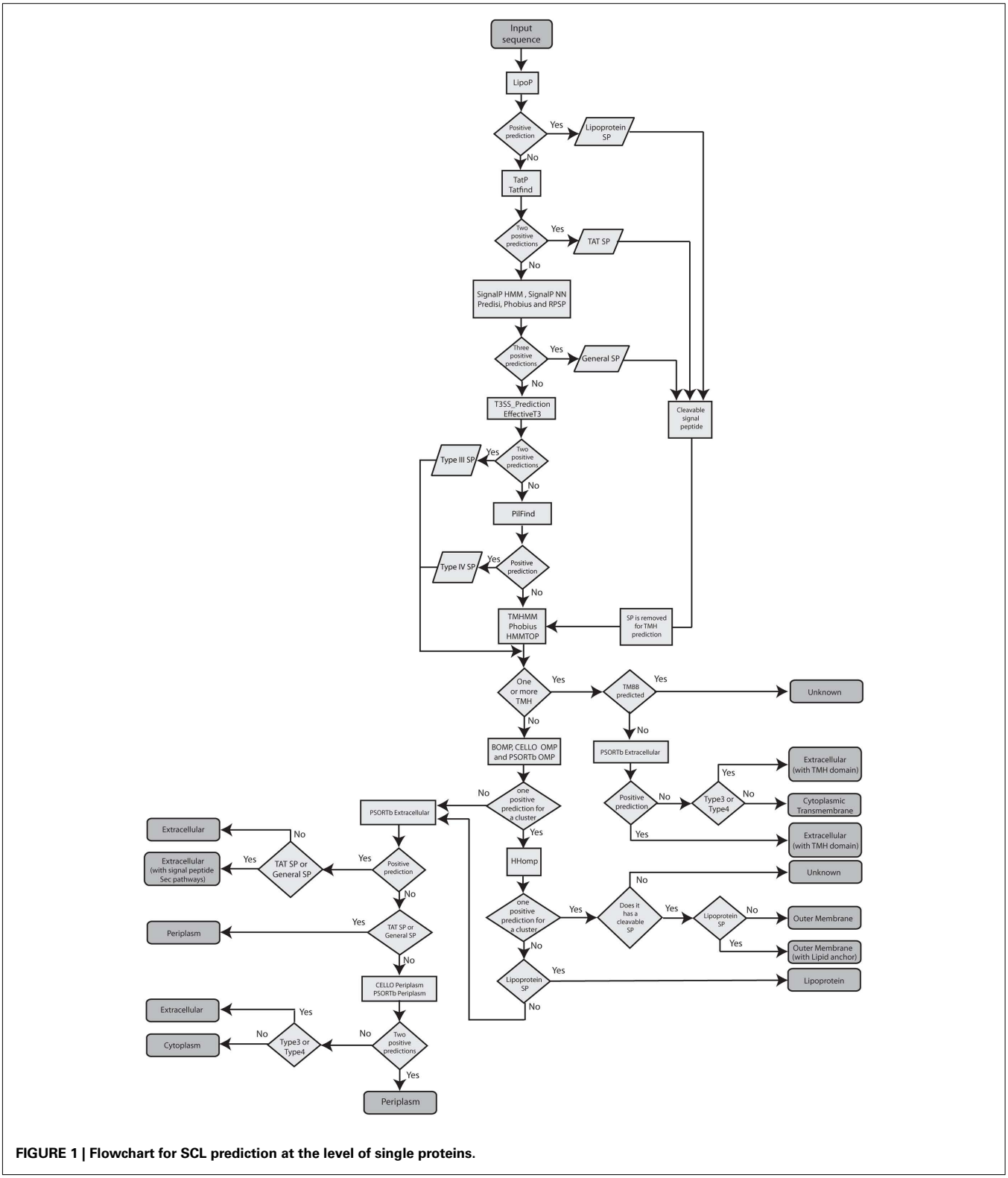


FIGURE 1 | Flowchart for SCL prediction at the level of single proteins.

Subcellular localization on the level of sequence clusters

To add homology information to single-sequence results in order to improve the overall prediction quality, all protein sequences from the DB_ClubSub-P dataset were clustered using CD-HIT (Li and Godzik, 2006); the clustering parameters are given in the Section "Datasets," above.

Since we cannot infer homology from singletons, we skipped 291,727 singletons and used the remaining 1,620,033 sequences,

which resulted in 174,028 clusters with sequence numbers ranging from 2 to 1,667. If a fraction of 0.7 or above of all proteins in the cluster have the same given SCL (i.e., 70% or more), this SCL is considered the SCL of the respective cluster. Clusters where no single SCL amounts to a protein fraction ≥0.7 (including "unknown") were annotated as "uncertain," and details of the predictions are kept available in the database for expert users to study further. Note that "uncertain" clusters are different from "unknown" clusters, as in the "unknown" ones most of the sequences show contradictory predictions to the rules described in the above section. Dual localization annotations were allowed only when two SCLs amounted to a fraction ≥0.7. The cutoff of 0.7 was chosen because any higher cutoff value leads to a steep increase in the number of "uncertain" clusters (see **Figure 3**).

SUBCELLULAR LOCALIZATION PREDICTION FOR ARCHAEA

We created a similar protocol to expand ClubSub-P to archaeal proteins. To this end, we used proteins from 65 archaeal proteomes (shown in Data Sheet S1 in Supplementary Material). After removing 779 small proteins with length 40 and below, we obtained 151,553 proteins for clustering using CD-HIT (Li and Godzik, 2006) with the same parameters as above. This resulted in 22,184 clusters with cluster size two and above. We named this dataset as DB_ClubSub-P_Archaea.

We used a similar parser to obtain a test dataset for Archaea. We obtained all the reviewed archaeal sequences without any "potential," "by similarity," or "probable" annotations in their SCL. We thus obtained 744 archaeal sequences with SCL annotation from UniProt Release 2011_07 (UniProt-Consortium, 2010). Sequences with ≤40 aa length were removed from the dataset and a non-redundant dataset with 40% sequence identity was created using CD-HIT (Li and Godzik, 2006), resulting in 252 sequences for the performance test. We named this dataset DB_experimental_Archaea.

For archaeal proteins, Lipoprotein, and TAT signal peptides were predicted using the same tools (LipoP, TatP, TatFind) as for Gram-negative bacterial proteins. For general signal peptide prediction, SignalP in Gram-positive mode was used, and Predisi was replaced by the tool PRED-SIGNAL (Bagos et al., 2009), which is an archaeal signal peptide prediction program. Phobius was used in default mode for the predictions. FlaFind (Szabó et al., 2007) was used to predict archaeal prepilin signal peptides; here, a TMH follows the signal peptide, and Prepilin peptidase cleaves the signal peptide before the TMH (Szabó et al., 2007). Thus, the protein is anchored to the membrane.

When two or more SPs were predicted, a consensus SP was annotated using a similar hierarchy as described in **Figure 1**, with the exception that there is no T3SS SP prediction for Archaea. Consensus TMH prediction was performed the same way as for Gram-negative bacteria. Archaeal proteins with TAT, general, or prepilin signal peptides or with PSORTb extracellular predictions (Yu et al., 2010) were annotated as "secreted/extracellular." Proteins with lipoprotein SP were annotated as "lipoproteins." "Cell wall" binding proteins were predicted using PSORTb's cell wall predictions (Yu et al., 2010). Proteins with one or more consensus TMH prediction were annotated as "cytoplasmic membrane" proteins. Proteins without any membrane domains or signal peptides or cell wall annotations were annotated as "cytoplasmic"

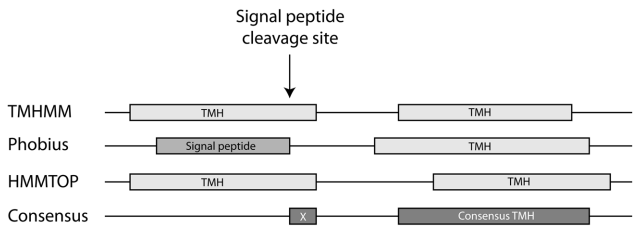


FIGURE 2 | Consensus transmembrane helix prediction module. A consensus TMH should be predicted by at least two tools. Signal peptides frequently result in false-positive TMH predictions and are removed with consensus TMH predictions.

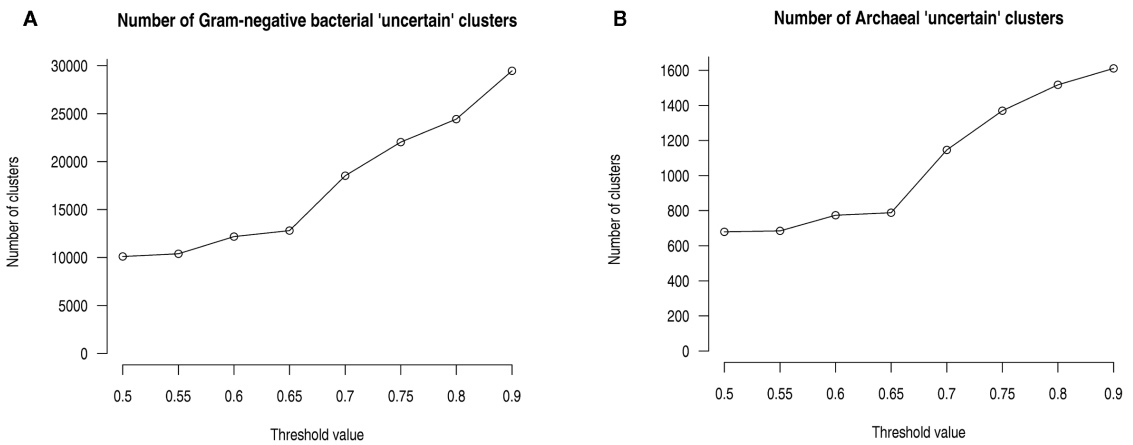


FIGURE 3 | Threshold determination for the assignment of subcellular localizations to whole clusters. (A) For Gram-negative bacteria; **(B)** for Archaea. In both cases, the number of clusters annotated as "uncertain" (i.e., with no SCL prediction above the threshold) increases at 0.7 – to minimize the number of "uncertain" clusters, cluster SCLs were assigned when a fraction of 0.7 or above (=70%) of the proteins in a cluster are predicted to have a given SCL.

Table 3 | Logical rules used for archaeal SCL predictions.

Features	Lipoprotein SP	TAT SP	General SP	Prepilin SP	Consensus TMH	PSORTb cell wall	PSORTb extracellular
LOCALIZATION							
Cytoplasm	No	No	No	No	No	No	No
Cytoplasmic membrane	Yes or no				One or more	Yes or no	Yes or no
Cell Wall	No	Yes or no			0 or more	Yes	No
Secreted/extracellular	No	Any one of the SP			0 or more	No	Yes or no

proteins. **Table 3** explains the rules for SCL prediction for Archaea.

DATABASE

We built a database from the above SCL annotations, which we named ClubSub-P, for "Cluster-based Subcellular localization Prediction." Results and input features are stored in SQL tables. The database is integrated into the classification section of the MPI Bioinformatics Toolkit (Biegert et al., 2006). The database is fully searchable using keywords or GI identifiers; moreover, FASTA sequences can be entered and will be assigned to the appropriate cluster through an internal BLAST search at >75% sequence coverage and >40% identity cutoff.

EVALUATION

We used the previously described DB_Experimental datasets to compare the performance of ClubSub-P with state-of-the-art SCL prediction tools. We calculated the precision, recall, accuracy, and the Mathew's correlation coefficient (MCC) for performance measure. In the following equations TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

Precision is a measure of the ability of the system to predict only the relevant data and it was calculated as the ratio between the number of predicted true positives against all positively predicted values, $TP/(TP + FP)$.

Recall is a measure of the ability of the system to predict all the relevant data and was calculated as the ratio between the number of predicted true positives against all true values, $TP/(TP + FN)$.

The accuracy of the system is defined by the closeness of its prediction toward the true values and was calculated by $(TP + TN)/(TP + TN + FP + FN)$.

The MCC calculates the correlation between the prediction and the observation and was calculated by $(TP * TN) - (FP * FN)/\sqrt{((TP + FN) * (TP * FP) * (TN + FP) * (TN + FN))}$.

RESULTS

CLUSTERING USING THE PSORTb v3 GRAM-NEGATIVE BACTERIAL TRAINING DATASET

As a first step, we had to make sure that the transfer of SCL information between homologous proteins is legitimate, and at which cutoffs for clustering (sequence identity and sequence coverage) this is still a valid procedure. To this end, we tested various clustering parameters using the 8,227 sequences in the DB_ePSORT dataset at decreasing cutoffs. To avoid problems with multi-domain proteins that might have different functions, and thus SCL, we decided to keep high sequence coverage. At 40% sequence identity and 80% sequence coverage, 6,136 sequences of the test set were clustered into 1,023 clusters with at least two sequences. 964 (94.2%) of these clusters had one common SCL for all of the proteins in the cluster. 47 (4.6%) of the clusters contained proteins with multiple SCL annotations which partially overlapped, and only 12 (1.2%) of the clusters had proteins with contradictory SCLs in them. Consequently, clustering done with the same parameters on the DB_ClubSub-P dataset can be expected to a have high number of clusters with homologous sequences that have a common SCL.

There are reports of orthologous proteins that have different SCLs in different organisms as a result of different evolutionary requirements. One prominent example is the glycerophosphoryl diester phosphodiesterase GlpQ, which is a periplasmic enzyme in *E. coli*, but is a surface-exposed lipoprotein in *Haemophilus influenzae* (Protein D; Janson et al., 1992). Such cases are rare, but they are easily missed when inferring their SCL from homology alone. In the case of Protein D/GlpQ, the two proteins are correctly predicted to have their respective – and different – SCL. Thus, one should always have a close look at the single-protein SCLs in cases where clustering leads to unclear or contradictory localization information. The ClubSub-P database allows for such manual inspection.

CLUSTER-BASED COMPARISON OF SIGNAL PEPTIDE AND TRANSMEMBRANE PREDICTION TOOLS

Applying a feature prediction tool such as a signal peptide predictor to sequences in a cluster of orthologous proteins should return similar results for all the proteins in the cluster (with very few but notable exceptions, see above). Inconsistency in such predictions will most probably be due to a lack of precision of the respective tool. Since different tools already use most of the proteins with experimentally verified SCL in their training sets, examining the performance of a tool at the cluster level is better suited to measure its sensitivity in a larger dataset, and to compare different tools.

Figure 4 shows the performance of different signal peptide prediction tools on our clusters produced from the DB_ClubSub-P dataset. We used only clusters with more than four sequences in this analysis; in detail, 66,716 clusters were included containing 1,341,180 sequences. If only <20% of the sequences in a cluster were positively predicted to contain a signal peptide, these predictions were assumed to be false positives; in a cluster with >80% of the sequences positively predicted, the remaining differing sequences were assumed to be false negatives.

Using these assumptions, we compared the TAT, type III, general signal peptide, and IM helix prediction tools along with their consensus predictions. Positive predictions from both TatP (Bendtsen et al., 2005) and TaTFind (Rose et al., 2002) tools

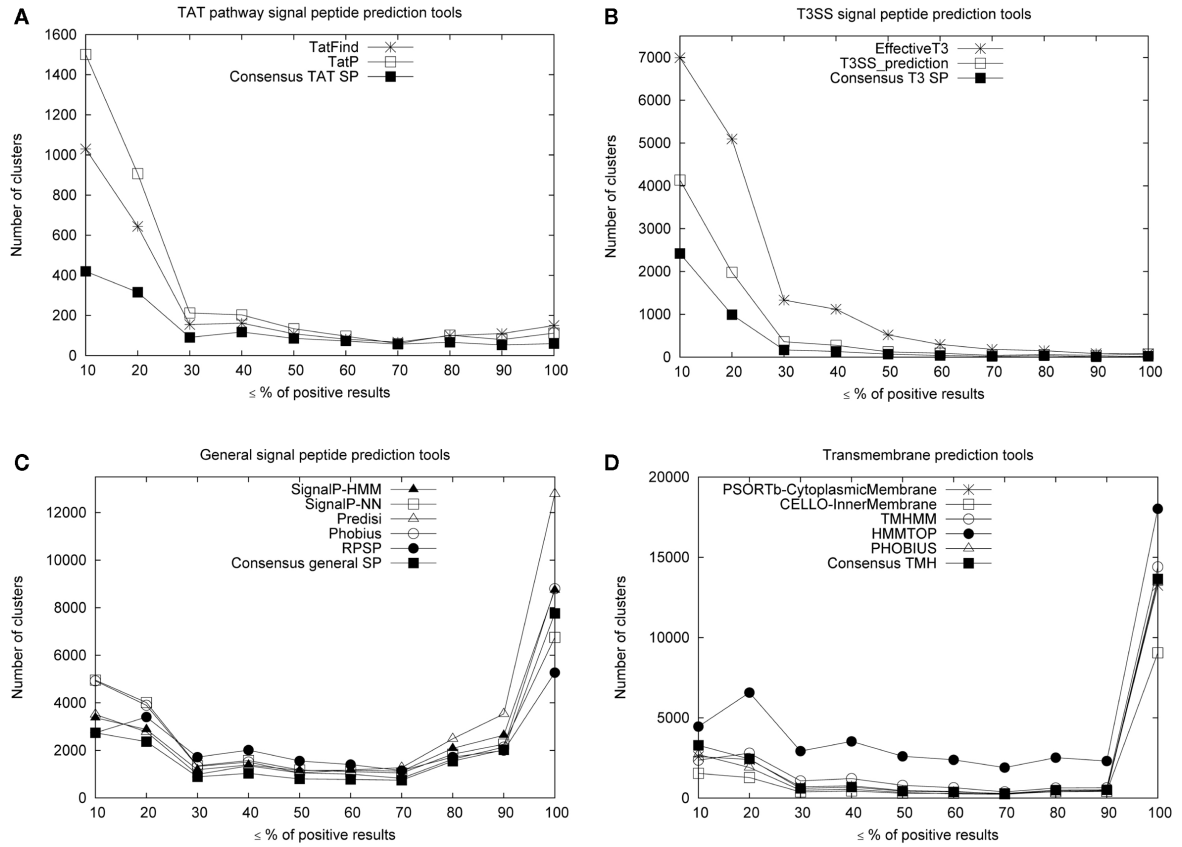


FIGURE 4 | Cluster-based comparison of signal peptide prediction tools. Shown are the comparisons for TAT **(A)**, T3SS **(B)**, General signal peptide **(C)** and transmembrane **(D)** prediction tools. The *x*-axis describes how many sequences (in %) in a cluster are predicted to have a signal peptide, where 10% means 0.1–10%, 20 means 10.1–20% etc. The majority of clusters contains only sequences where no signal sequence is predicted (0% positive results) – for clarity, these are ignored in the graph.

were considered as a consensus TAT signal peptide. False-positive predictions were largely reduced by these consensus predictions (**Figure 4A**). This shows that most of the positive predictions in clusters with <20% positives are in fact false-positive predictions from the tools. A similar result can be seen with the consensus prediction for type III signal peptides (**Figure 4B**), where we considered positive predictions from both T3SS_prediction (Löwer and Schneider, 2009) and EffectiveT3 (Arnold et al., 2009) tools as a consensus type III signal peptide. For consensus general signal peptide prediction, we required at least three positive predictions from highly precise general signal peptide tools (Choo et al., 2009) like SignalP-HMM, SignalP-NN (Bendtsen et al., 2004), Predisi (Hiller et al., 2004), RPSP (Plewczynskia et al., 2007), and Phobius (Käll et al., 2004). **Figure 4C** shows the cluster-based comparison for general signal peptide tools and the consensus made from their prediction.

Similarly, we compared the performance of TMH prediction by CELLO v.2.5 (Yu et al., 2006), PSORTb v3.0.2 (Yu et al., 2010), Phobius (Käll et al., 2004), TMHMM 2.0 (Krogh et al., 2001), and HMMTOP v2.0 (Tusnady and Simon, 2001; **Figure 4D**), assuming that prediction of at least one TMH indicates that the protein is a transmembrane protein. The result clearly shows the high false-positive rate of HMMTOP predictions (**Figure 4D**), compared to the predictions of the other tools. However, the consensus TMH prediction of Phobius, TMHMM 2.0, and HMMTOP v2.0 (see Materials and Methods) eliminated most of these false-positive predictions.

Such comparisons of different prediction tools help in selecting the best tools for consensus predictions; alternatively, one could use this performance measure to weigh different tools, giving more importance to tools that performed better.

CLUBSUB-P DATABASE STATISTICS

The core of the cluster-based SCL prediction is the ClubSub-P database. Of 2,331,935 retrieved sequences, 404,542 identical sequences and 15,633 sequences with less than 40 residues and were removed. The remaining 1,911,760 sequences were clustered using CD-HIT (Li and Godzik, 2006). We used 40% local sequence identity at 80% sequence coverage for clustering, When these settings were applied, 1,620,033 sequences (84.74%) were clustered into 174,028 clusters with size range from 2 to 1,677 and 291,727 proteins (15.26%) appeared to be singletons, meaning these sequences do not have any homolog among the sequences in the database at these settings. These singletons were not analyzed in detail, since no homology information can be inferred for them. Note though that with expansion of the database, these

proteins might fall into newly formed clusters at a later time point as discussed below.

We were able to annotate the SCL of 1,500,778 of 1,620,033 sequences that are grouped in clusters of at least two sequences, which is 78.50% of the sequences used in clustering (1,911,760 sequences – note again that singletons, i.e., sequences that do not fall into clusters, are excluded from our predictions). For comparison, PSORTb v3.0.2 annotates 71.25% of all sequences used in our clustering approach (1,362,110 of 1,911,760 sequences). The details of the ClubSub-P prediction statistics for Gram-negative bacteria are shown in **Table 4**.

MULTIPLE SUBCELLULAR LOCALIZATION PREDICTIONS

In addition to the common SCL classifications in Gram-negative bacteria, we found clusters of proteins with features that correspond to two different SCLs, e.g., "extracellular" proteins that have signal peptides for secretion to the "periplasm," "extracellular" proteins with "TMHs" to get inserted into host membranes, and "OM β-barrel" proteins with a "lipoprotein" signal peptide. In many cases, experimental evidence for these double localizations exists, demonstrating that they are not artifacts of our SCL prediction pipeline. As an example, the "Pertussis toxin subunit 1" (UniProt ID – TOX1_BORPE/gi|33594638) is predicted by ClubSub-P to have an "extracellular" and a "periplasmic" localization; by experimental evidence (Farizo et al., 2002) it is an extracellular protein that is first secreted to the periplasm using the general signal peptide pathway, and only subsequently is secreted to the extracellular space. Moreover, the "Outer membrane protein oprM" (UniProt ID – OPRM_PSEAE/gi|116054158; Nakajima et al., 2000) has been shown experimentally to be attached to the OM via a lipid anchor, while it also spans the OM with a TMBB domain. ClubSub-P predicts OprM to be an OM β-barrel protein as well as a lipoprotein. A prominent example for "extracellular" and "transmembrane" localization are proteins secreted by pathogens to insert in to the host membrane, such as the needle tip components of the Type III secretion apparatus (Marlovits and Stebbins, 2010); and indeed, we find SipB from *Salmonella* (UniProt ID – SIPB_SALTY/gi|62181387) among the proteins with both extracellular and transmembrane localization. Thus, double localizations in our database, while sometimes counterintuitive, can reflect important information on complex secretion pathways.

Table 4 | Statistics of the ClubSub-P database.

ClubSub-P subcellular localizations	No. of clusters	No. of proteins
Cytoplasmic	95,191	1,023,339
Cytoplasmic membrane	33,814	304,996
Periplasmic	15,261	107,602
Inner/outer membrane lipoprotein	4,471	27,711
Outer membrane beta-barrel	3,011	20,976
Extracellular	1,319	8,250
Extracellular AND transmembrane helix	733	3,582
Extracellular AND signal peptide	540	2,930
Outer membrane beta-barrel AND lipid anchor	124	1,572
Uncertain[1]	18,388	113,286
Unknown[2]	1,356	5,969

[1] *Uncertain are the clusters where none of the SCLs, including "unknown," are above the 70% threshold.*

[2] *Unknown are the clusters where "unknown" SCL was above the threshold of 70%. This is usually due to contradictory SCL predictions.*

PERFORMANCE MEASURE

The performance of ClubSub-P was compared to PSORTb v.3.0.2 (Yu et al., 2010) and CELLO v.2.5 (Yu et al., 2006). We calculated the precision, recall, accuracy, and MCC.

Unfortunately, the Proteome Analyst prediction server (Lu et al., 2004) is not active any more, thus we could not compare ClubSub-P against it. A recently published database for SCL prediction of Gram-negative bacteria, CobaltDB v1 (Goudenège et al., 2010), provides meta predictions for different signal peptide and secondary structural features; however, it does not combine these results to annotate a final SCL for the proteins. For this reason we could not use CobaltDB in our performance measure.

Dual localization predictions were considered for all the tools compared in the performance measure, but only CELLO and the UniProt original annotations had proteins with dual annotations in our test dataset. However, proteins with more than two localization predictions in CELLO v.2.5 were not considered and annotated as unknown. In cases where two different SCLs for a single protein are either predicted by a tool or given from UniProt data in the test set, a hit is considered as "true positive" if at least one of the localizations matches. All the "Unknown" predictions were considered as false negatives in our performance measurements. Sequences from test datasets were used to search against the ClubSub-P database, in order to assign their SCL. Only hits with sequence identities above 40% and pairwise alignment coverage above 75% were annotated to the corresponding cluster and sequences with no hits or below this cutoff were assigned as "Unknown". The hits with "Uncertain" localization (see Materials and Methods) were also considered as "Unknown" for the performance measurements.

The results of the performance measurement are shown in **Table 5**. With the DB_Experimental test dataset, ClubSub-P (83.85%) shows a higher precision than PSORTb v3.0.2 (80%) and CELLO v2.5 (66.67%). Since the recall value for periplasmic proteins (15.79%) is very low for PSORTb v3.0.2, the overall recall value of PSORTbv3.0.2 (54.55%) is lower than that of CELLO (70.18%) and ClubSub-P (62.64%). Overall, the accuracy of all tools is comparable. Since we considered any one of correct dual localization predictions as "true positive," CELLO's overall performance (0.6) in terms of MCC is comparable to PSORTb (0.59). ClubSub-P has a superior overall performance (MCC 0.67). In summary, ClubSub-P has a higher precision than PSORTb and CELLO, showing that its strength is a reduced false-positive rate through the use of homology information.

INCORRECT START CODONS RESULTING IN MISANNOTATED SIGNAL PEPTIDES

A known problem in SCL prediction is the quality of the input sequences; especially the exact start position for proteins with N-terminal signal peptides is essential. In the course of

our analysis, we noted that in clusters where the majority of sequences are predicted to contain an N-terminal signal peptide, the false-negative results typically stem from misannotated start codons. When we corrected such gene annotation errors in the sequence, the signal peptides were correctly predicted in most cases. We found examples for both possible cases, where the misannotated start codons either extended or shortened the sequence N-terminally. Examples for these cases are shown in **Figure 5**; the sequences in this cluster that contained misannotated start codons were not predicted to contain a signal peptide, but had a OM beta-barrel annotation and thus were annotated as unknown, while the correctly annotated sequences in the cluster were predicted to be OM beta-barrel protein with a lipid anchor. The SCL annotation of the cluster reassigns them to OMP proteins with lipid anchor via the cluster consensus annotation, which shows one strength of ClubSub-P, the additional use of homology information on top of single-sequence predictions.

Overall, we found 3,558 proteins with false-negative predictions in different clusters of proteins with signal peptides (annotated as periplasmic, OMP, OMP with lipid anchor, lipoproteins, or extracellular with signal peptide). These 3,558 proteins were spread across 547 of the 607 genomes that we used in this study, and were present in 2,222 different clusters (Data Sheet S2 in Supplementary Material). These errors were significantly accumulated in certain genomes compared with the rest of the genomes in the database (**Table 6**). This could be due to differences in the gene prediction and ORF finder methods used in the gene annotation process. But as we can easily find these mistakes only in the signal peptide-containing clusters, we cannot provide good statistical data on the performance of the different gene prediction pipelines – there might be additional misannotations in other proteins that do not have an N-terminal signal peptide.

Table 5 | Performance measurement for different Gram-negative bacterial subcellular localization prediction tools.

Location	Precision	Recall	Accuracy	MCC
PSORTbv3				
Cytoplasm	66.67	74.42	83.93	0.6
Inner membrane	90	58.06	90.68	0.68
Periplasm	60	15.79	89.41	0.27
Outer membrane	55.56	62.5	95.88	0.57
Extracellular	100	50.67	78.24	0.6
Total	80	54.55	87.6	0.59
CELLO				
Cytoplasm	62.32	100	84.34	0.7
Inner membrane	94.12	61.54	92.95	0.73
Periplasm	58.62	89.47	91.3	0.68
Outer membrane	28.57	75	89.7	0.42
Extracellular	86.36	50.67	74.25	0.5
Total	66.67	70.18	86.38	0.6
CLUBSUB-P				
Cytoplasm	72.22	88.64	88.17	0.72
Inner membrane	100	53.57	91.77	0.7
Periplasm	73.68	73.68	94.12	0.7
Outer membrane	87.5	87.5	98.82	0.87
Extracellular	100	45.33	75.88	0.56
Total	83.85	62.64	89.73	0.67

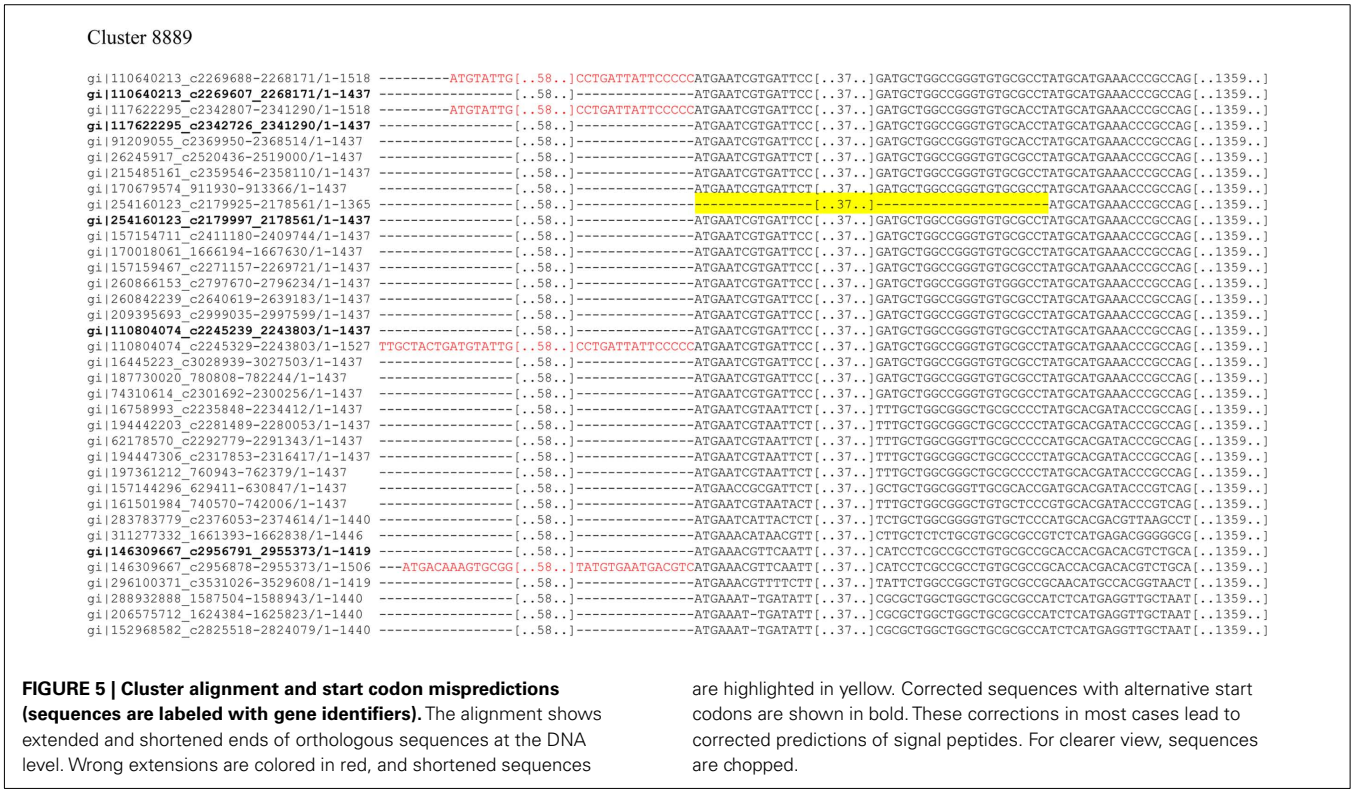


FIGURE 5 | Cluster alignment and start codon mispredictions (sequences are labeled with gene identifiers). The alignment shows extended and shortened ends of orthologous sequences at the DNA level. Wrong extensions are colored in red, and shortened sequences are highlighted in yellow. Corrected sequences with alternative start codons are shown in bold. These corrections in most cases lead to corrected predictions of signal peptides. For clearer view, sequences are chopped.

Random manual checking revealed that most of the 3,558 protein sequences with false-negative signal peptide predictions have a mispredicted start codon on the DNA level. Studies have shown that the biased use of the uncommon start codons GUG and UUG over AUG is common among mispredicted start codons (Starmer et al., 2006; Pallejà et al., 2008). Confirming these findings, we also found a biased use of uncommon start codons among the above mentioned 3,558 proteins. The frequency of start codon usage in all bacterial coding sequences (3,690,458 sequences) used for this analysis is AUG (80.7%), GUG (12.6%), UUG (6.5%), and other start codons (0.2%). But the gene start codon frequencies of the 3,558 falsely predicted proteins are AUG (62.73%), GUG (21.61%), UUG (12.45%), and other start codons (3.2%), again showing that these gene predictions need revision.

We wanted to check if we could detect signal peptides from the genes with alternative start codons after re-annotation. ProTISA (Hu et al., 2008) is a database which combines translation initiation site (TIS) information from different sources, e.g., from experimental Swiss-Prot annotations, conserved domain hits and from alignments of orthologous sequences, to refine the RefSeq TIS annotations. Unfortunately, it doesn't cover all the proteomes we used in our database; thus, we used the alternative start codons predicted by gene prediction programs instead (see above). The NCBI RefSeq FTP site provides updated gene predictions for all sequenced bacterial genomes, based on the latest version of four gene prediction programs [GeneMark-2.5m (Borodovsky and Mcininch, 1993), GeneMarkHMM-2.6r (Borodovsky and Lukashin, 1998), Glimmer3 (Delcher et al., 2007), and Prodigal-2.50 (Hyatt et al., 2010)]. To obtain more quantitative information on the phenomenon, we used this precomputed data to find an alternative start codon for the 3,558 proteins with false-negative signal peptide predictions (see methods), which translates into a protein with a signal peptide according to SignalP-HMM. Together, 2,290 sequences with an alternative start leading to a positive signal peptide prediction were found by one or several gene prediction programs. Of these 2,290 positive predictions, GeneMark-2.5m predicts 69.91% (1,601), GeneMarkHMM-2.6r predicts 72.79% (1,667), Glimmer3 predicts 66.86% (1,531), and Prodigal-2.50 predicts 84.93% (1,945). The numbers do not significantly change by using LipoP or Phobius instead of SignalP-HMM. The details of the alternative start codons with positive signal peptide predictions are given in Data Sheet S2 in Supplementary Material.

Table 6 | Genomes with multiple signal peptide/start codon errors in secretory clusters.

Replicon name	Number of alternative start codons*	Replicon ID
Acinetobacter baumannii ATCC 17978	104	NC_009085
Cronobacter turicensis z3032	62	NC_013282
Pseudomonas putida S16 chromosome	27	NC_015733
Shewanella violacea DSS12 chromosome	27	NC_014012
Caulobacter crescentus CB15 chromosome	25	NC_002696
Klebsiella pneumoniae subsp. *pneumoniae* MGH 78578 chromosome	21	NC_009648
Shewanella piezotolerans WP3 chromosome	20	NC_011566

**Found in protein clusters with signal peptide annotation where single-sequences lacked the signal peptide. Only genomes with more than 20 erroneous proteins are shown.*

SUBCELLULAR LOCALIZATION IN ARCHAEA

Archaea have a comparable cellular architecture to Gram-positive bacteria, except that instead of a peptidoglycan layer, different types of surface layers made from proteins, glycoproteins, or pseudo-murein are observed (Ellen et al., 2010). As there are no specialized SCL prediction programs available for Archaea other than the recently published PSORTb v3.0.2 program (Yu et al., 2010), we combined different feature prediction tools along with homology information in the same way as described above for Gram-negative bacteria.

As a result we were able to assign unambiguous SCLs to 69.21% of all proteins obtained from the 65 archaeal proteomes (104,896 of 151,553), where PSORTbV3.0.2 annotates 86.99% (131,839 of 151,553). When exclusively looking at proteins found in clusters with size two and above, i.e., where homology information is available, ClubSub-P can annotate 96.35% (104,896 out of 108,872) of proteins with an unambiguous SCL, where PSORTbv3.0.2 predicts only 89.42% (97,349 of 108,872).

ClubSub-P archaeal SCL annotation statistics are found in the **Table 7**. Just like in the Gram-negative SCL predictions, we also found clusters of archaeal proteins with multiple localizations, such as "Secreted/extracellular AND membrane anchor" and "Cell wall AND membrane anchor." We annotated these combinations separately as we assume that, again as for Gram-negative bacteria, these double localizations have a biological significance. In detail, proteins with a predicted signal peptide and one consensus membrane helix prediction were annotated as "Secreted/extracellular AND membrane anchor." This also includes the proteins with prepilin signal peptide. Proteins with a cell wall prediction and one or two consensus membrane

Table 7 | ClubSub-P archaeal SCL prediction statistics.

Cluster's subcellular localizations	No. of clusters	No. of sequences
Cytoplasmic	15,592	84,978
Cytoplasmic membrane	4,535	17,158
Secreted/extracellular	399	1,157
Secreted/extracellular with membrane anchor	244	804
Lipoprotein	181	572
Cell wall	57	189
Cell wall with membrane anchor	14	38
Uncertain	1,139	3,921
Unknown	23	55

helix predictions were annotated as "Cell wall AND membrane anchor." Note that membrane anchor in this context means a single N-terminal transmembrane helix that anchors proteins to the cytoplasmic membrane.

Since there are very few archaeal proteins with experimentally annotated SCLs, most of these proteins are already included in the training sets of the tools we used in the consensus prediction, which makes the calculation of benchmarks very difficult. But, using the experimentally verified 252 archaeal sequences from UniProt, we were able to show that ClubSub-P has a slightly higher precision than PSORTbV3.0, but with a lower recall value. Overall, both tools are comparable in performance. Details on the performance of ClubSub-P with archaeal proteins are found in **Table 8**. With the addition of more archaeal proteomes from genomic data, and with the inclusion of further tools specialized on SCL prediction of archaeal proteins, ClubSub-P will be able to predict the archaeal SCLs more precisely in the future.

CLUBSUB-P AVAILABILITY

We introduced the ClubSub-P database into the classification section of the MPI Bioinformatics Toolkit, a platform that integrates a great variety of tools for protein sequence analysis (Biegert et al., 2006). ClubSub-P can be found at http://toolkit.tuebingen.mpg.de/clubsubp. Users can browse the database to view the precomputed results, or they can annotate their query sequences by searching the database using BLAST.

Table 8 | Performance measurement of ClubSub-P archaeal predictions.

	Precision	Recall	Accuracy	MCC
PSORTb v3.0.2	98.8	98.02	99.2	0.98
ClubSub-P	99.55	86.77	96.46	0.91

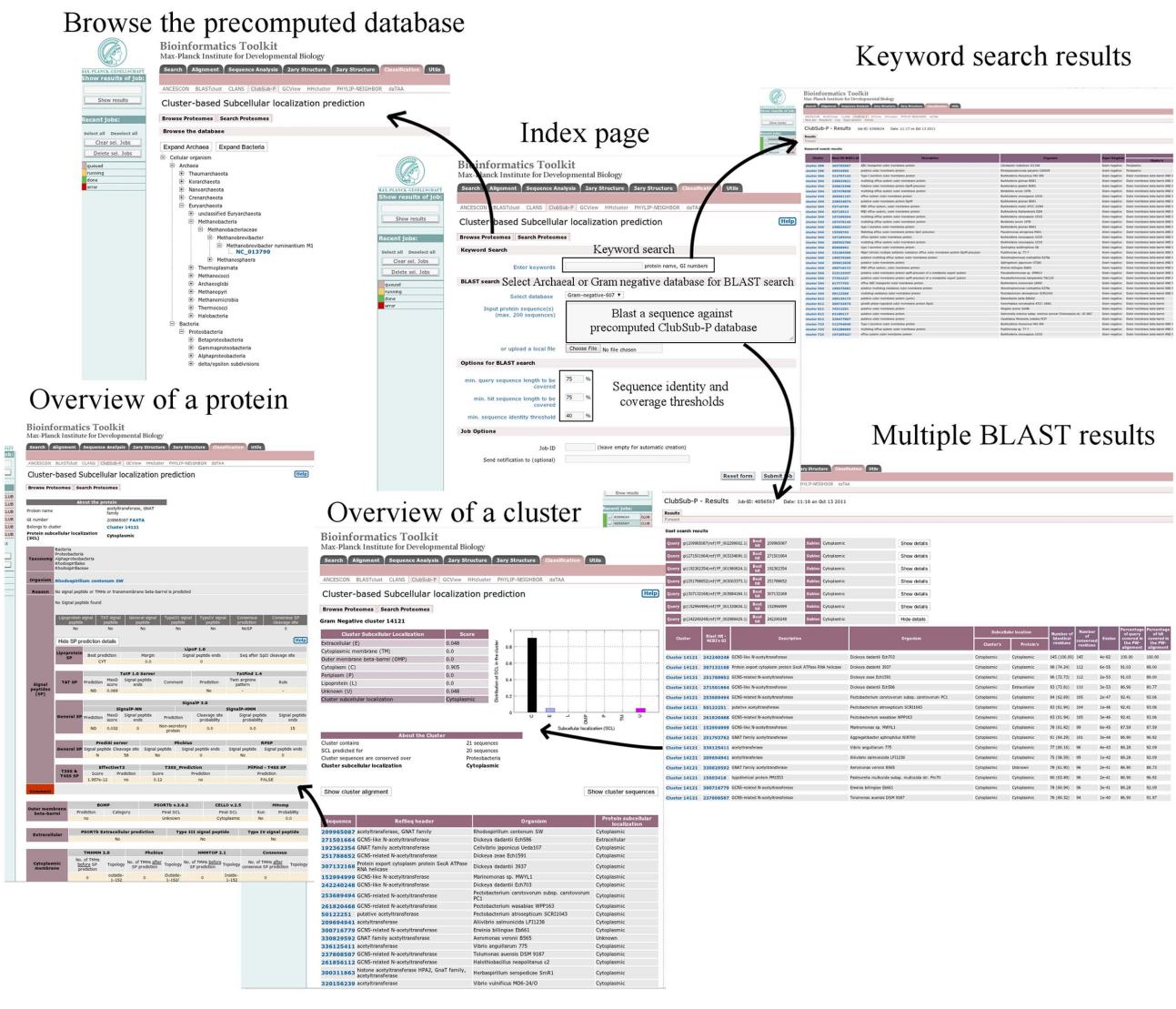


FIGURE 6 | ClubSub-P database screenshots.

ClubSub-P is interconnected with other tools in the toolkit, so users can easily forward their results to other tools for further analysis. Screenshots of the ClubSub-P database are shown in **Figure 6**.

DISCUSSION

Annotating the SCL of a protein is an important step in characterizing the native function of a protein. Thus, computational SCL predictions have gained importance in the post-genomic era, and various tools exist for this purpose. When combing different SCL predictors to create a meta-SCL predictor, it is important to select the best available individual predictors. We have developed a cluster-based meta-SCL prediction method for archaeal and Gram-negative bacterial proteins, by combining different published tools through consensus voting and protein sorting rules. In addition to the consensus SCL prediction for each single sequence, sequences are clustered according to their similarity. This homology information is exploited to eliminate false-positive and false-negative results. The performance of our tool is comparable with state-of-the-art SCL prediction methods, but with more precision (where precision is a measure of the ability of the system to predict only the relevant data, see Materials and Methods). In addition to the general SCLs, we were able to annotate more specific localizations, such as "OMP with lipid anchor," "extracellular protein with transmembrane helix," and "transmembrane with TAT or general signal peptide" for certain protein clusters, by combining different feature prediction tools. When more of such specific feature prediction tools become available we can include them into our prediction pipeline easily, and can annotate more specific localizations in a very precise way. In the cluster-based comparison of predictions for orthologous proteins, we have shown that there are inconsistencies between different prediction methods. We have demonstrated that by obtaining a consensus prediction from different tools, we can greatly reduce the number of false-positive predictions for single sequences. Furthermore, combining the single SCL predictions on the level of clusters further increases the precision of the predictions. The incorporation of additional proteomes from new sequencing projects will further decrease the number of singletons and will significantly increase the coverage and the precision of the SCL predictions of ClubSub-P in the future.

The pipeline can be expanded to other organism groups easily, as we show with the example of Archaea. Archaea are especially interesting in this context as comparably little experimental information is available for them. As only few reliable SCL prediction tools trained specifically on archaeal datasets are available, ClubSub-P is at an advantage as it combines different tools into a (more reliable) consensus prediction, and uses homology information where available to exclude most false-positive and false-negative predictions. Though the recall value is lower than that of PSORTb, the overall performance will increase dramatically by adding more sequenced archaeal genomes for clustering, and with new and Archaea-specific SCL prediction tools which can be incorporated into ClubSub-P easily.

The database can be used for a variety of applications. One obvious application is in genome annotation, where we show how misinterpreted start codons can be detected through SCL predictions and the use of homology information. We originally produced the database to screen for conserved immunogenic epitopes localized on the bacterial cell surface, in order to identify new vaccine candidates which would protect from diseases caused by Gram-negative human pathogens. Using the protein clusters with OM or extracellular localization, one can find conserved proteins which could be useful vaccine candidates or diagnosis markers specific to the bacterial species present in these clusters.

ACKNOWLEDGMENTS

The authors are very thankful to Christina Wassermann and Andre Noll for the useful discussions and technical assistance in integrating the database into the MPI toolkit, and to Vikram Alva, Thomas Arnold, Stanislaw Dunin-Horkawicz, Iwan Grin, Marcus Thein and others for helpful discussions, and to Andrei Lupas for continuing support. We are also thankful to the many authors that provided us with offline version of their tools. This work was funded by the Bill & Melinda Gates Foundation, Grand Challenges Explorations program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/evolutionary_and_genomic_microbiology/10.3389/fmicb.2011.00218/abstract

REFERENCES

Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., Niinikoski, A., Mewes, H.-W., Horn, M., and Rattei, T. (2009). Sequence-based prediction of type III secreted proteins. *PLoS Pathog.* 5, e1000376. doi:10.1371/journal.ppat.1000376

Bagos, P. G., Tsirigos, K. D., Plessas, S. K., Liakopoulos, T. D., and Hamodrakas, S. J. (2009). Prediction of signal peptides in Archaea. *Protein Eng. Des. Sel.* 22, 27–35.

Bendtsen, J., Nielsen, H., Widdick, D., Palmer, T., and Brunak, S. (2005). Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 6, 167. doi:10.1186/1471-2105-6-167

Bendtsen, J. D., Nielsen, H., Von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: signalP 3.0. *J. Mol. Biol.* 340, 783–795.

Berven, F. S., Flikka, K., Jensen, H. B., and Eidhammer, I. (2004). BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res.* 32, W394–W399.

Biegert, A., Mayer, C., Remmert, M., Söding, J., and Lupas, A. N. (2006). The MPI Bioinformatics Toolkit for protein sequence analysis. *Nucleic Acids Res.* 34, W335–W339.

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

Borodovsky, M., and Mcininch, J. (1993). GenMark: parallel gene recognition for both DNA strands. *Comput. Chem.* 17, 123–133.

Bos, M. P., Robert, V., and Tommassen, J. (2007). Biogenesis of the gram-negative bacterial outer membrane. *Annu. Rev. Microbiol.* 61, 191–214.

Burstein, D., Zusman, T., Degtyar, E., Viner, R., Segal, G., and Pupko, T. (2009). Genome-scale identification of *Legionella pneumophila* effectors using a machine learning approach. *PLoS Pathog.* 5, e1000508. doi:10.1371/journal.ppat.1000508

Choo, K., Tan, T., and Ranganathan, S. (2009). A comprehensive assessment of N-terminal signal peptides prediction methods. *BMC Bioinformatics* 10, S2. doi:10.1186/1471-2105-10-S15-S2

Chou, K.-C., and Shen, H.-B. (2006a). Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* 347, 150–157.

Chou, K.-C., and Shen, H.-B. (2006b). Large-scale predictions of gram-negative bacterial protein subcellular locations. *J. Proteome Res.* 5, 3420–3428.

Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23, 673–679.

Desvaux, M., Hébraud, M., Talon, R., and Henderson, I. R. (2009). Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol.* 17, 139–145.

Ellen, A. F., Zolghadr, B., Driessen, A. M., and Albers, S. V. (2010). Shaping the archaeal cell envelope. *Archaea* 2010, 608243.

Farizo, K. M., Fiddner, S., Cheung, A. M., and Burns, D. L. (2002). Membrane localization of the S1 subunit of pertussis toxin in *Bordetella pertussis* and implications for pertussis toxin secretion. *Infect. Immun.* 70, 1193–1201.

Gardy, J. L., and Brinkman, F. S. L. (2006). Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 4, 741–751.

Giombini, E., Orsini, M., Carrabino, D., and Tramontano, A. (2010). An automatic method for identifying surface proteins in bacteria: {SLEP}. *BMC Bioinformatics* 11, 39. doi:10.1186/1471-2105-11-39

Goudenège, D., Avner, S., Lucchetti-Miganeh, C., and Barloy-Hubler, F. (2010). CoBaltDB: complete bacterial and archaeal orfeomes subcellular localization database and associated resources. *BMC Microbiol.* 10, 88. doi:10.1186/1471-2180-10-88

Hiller, K., Grote, A., Scheer, M., Münch, R., and Jahn, D. (2004). PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* 32, W375–W379.

Horler, R. S. P., Butcher, A., Papangelopoulos, N., Ashton, P. D., and Thomas, G. H. (2009). EchoLOCATION: an in silico analysis of the subcellular locations of *Escherichia coli* proteins and comparison with experimentally derived locations. *Bioinformatics* 25, 163–166.

Hu, G.-Q., Zheng, X., Yang, Y.-F., Ortet, P., She, Z.-S., and Zhu, H. (2008). ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes. *Nucleic Acids Res.* 36, D114–D119.

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11, 119. doi:10.1186/1471-2105-11-119

Imai, K., and Nakai, K. (2010). Prediction of subcellular locations of proteins: where to proceed? *Proteomics* 10, 3970–3983.

Janson, H., Heden, L. O., and Forsgren, A. (1992). Protein D, the immunoglobulin D-binding protein of *Haemophilus influenzae,* is a lipoprotein. *Infect. Immun.* 60, 1336–1342.

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

Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 338, 1027–1036.

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

Lewenza, S., Mhlanga, M. M., and Pugsley, A. P. (2008). Novel inner membrane retention signals in *Pseudomonas aeruginosa* lipoproteins. *J. Bacteriol.* 190, 6119–6125.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659.

Löwer, M., and Schneider, G. (2009). Prediction of type III secretion signals in genomes of Gram-negative bacteria. *PLoS ONE* 4, e5917. doi:10.1371/journal.pone.0005917

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

Luirink, J., Von Heijne, G., Houben, E., and De Gier, J.-W. (2005). Biogenesis of inner membrane proteins in *Escherichia coli. Annu. Rev. Microbiol.* 59, 329–355.

Mah, N., Perez-Iratxeta, C., and Andrade-Navarro, M. A. (2010). Outer membrane pore protein prediction in mycobacteria using genomic comparison. *Microbiology* 156, 2506–2515.

Marlovits, T. C., and Stebbins, C. E. (2010). Type III secretion systems shape up as they ship out. *Curr. Opin. Microbiol.* 13, 47–52.

Nakajima, A., Sugimoto, Y., Yoneyama, H., and Nakae, T. (2000). Localization of the outer membrane subunit OprM of resistance-nodulation-cell division family multicomponent efflux pump in *Pseudomonas aeruginosa. J. Biol. Chem.* 275, 30064–30068.

Nielsen, H., Engelbrecht, J., Brunak, S., and Von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng. Des. Sel.* 10, 1–6.

Overbeek, R., Bartels, D., Vonstein, V., and Meyer, F. (2007). Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem. Rev.* 107, 3431–3447.

Pallejà, A., Harrington, E. D., and Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics* 9, 335. doi:10.1186/1471-2164-9-335

Plewczynski, D., Slabinski, L., Tkacz, A., Kajan, L., Holm, L., Ginalski, K., and Rychlewski, L. (2007). The RPSP: Web server for prediction of signal peptides. *Polymer* 48, 5493–5496.

Pugsley, A. P., Kornacker, M. G., and Ryter, A. (1990). Analysis of the subcellular location of pullulanase produced by *Escherichia coli* carrying the pulA gene from *Klebsiella pneumoniae* strain UNF5023. *Mol. Microbiol.* 4, 59–72.

Rashid, M., Saha, S., and Raghava, G. P. (2007). Support vector machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. *BMC Bioinformatics* 8, 337. doi:10.1186/1471-2105-8-337

Remmert, M., Linke, D., Lupas, A. N., and Söding, J. (2009). HHomp – prediction and classification of outer membrane proteins. *Nucleic Acids Res.* 37, W446–W451.

Rose, R. W., Bruser, T., Kissinger, J. C., and Pohlschroder, M. (2002). Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. *Mol. Microbiol.* 45, 943–950.

Saier, M. H., Ma, C. H., and Rodgers, L. (2008). Protein secretion and membrane insertion systems in bacteria and eukaryotic organelles. *Adv. Appl. Microbiol.* 65, 141–197.

Seydel, A., Gounon, P., and Pugsley, A. P. (1999). Testing the "+2 rule" for lipoprotein sorting in the *Escherichia coli* cell envelope with a new genetic selection. *Mol. Microbiol.* 34, 810–821.

Shen, Y. Q., and Burger, G. (2007). "Unite and conquer": enhanced prediction of protein subcellular localization by integrating multiple specialized tools. *BMC Bioinformatics* 8, 420. doi:10.1186/1471-2105-8-420

Starmer, J., Stomp, A., Vouk, M., and Bitzer, D. (2006). Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput. Biol.* 2, e57. doi:10.1371/journal.pcbi.0020057

Su, E. C.-Y., Chiu, H.-S., Lo, A., Hwang, J.-K., Sung, T.-Y., and Hsu, W.-L. (2007). Protein subcellular localization prediction based on compartment-specific features and structure conservation. *BMC Bioinformatics* 8, 330. doi:10.1186/1471-2105-8-330

Szabó, Z., Stahl, A. O., Albers, S.-V., Kissinger, J. C., Driessen, A. J. M., and Pohlschröder, M. (2007). Identification of diverse archaeal proteins with class III signal peptides cleaved by distinct archaeal prepilin peptidases. *J. Bacteriol.* 189, 772–778.

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

UniProt-Consortium. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142–D148.

Warren, A., Archuleta, J., Feng, W.-C., and Setubal, J. (2010). Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics* 11, 131. doi:10.1186/1471-2105-11-131

Yu, C.-S., Chen, Y.-C., Lu, C.-H., and Hwang, J.-K. (2006). Prediction of protein subcellular localization. *Proteins* 64, 643–651.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., and Brinkman, F. S. L. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608–1615.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 September 2011; accepted: 12 October 2011; published online: 08 November 2011.

*Citation: Paramasivam N and Linke D (2011) ClubSub-P: cluster-based subcellular localization prediction for Gram-negative bacteria and archaea. Front. Microbio. **2**:218. doi: 10.3389/fmicb.2011.00218*

This article was submitted to Frontiers in Evolutionary and Genomic Microbiology, a specialty of Frontiers in Microbiology.

Is the C-terminal insertional signal in Gram-negative bacterial outer membrane proteins species-specific or not?

Nagarajan Paramasivam[1], Michael Habeck[1], Dirk Linke[1]*

[1]Department I, Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany

*Corresponding author

Email address:

NP: nagarajan.paramasivam@tuebingen.mpg.de

MH: michael.habeck@tuebingen.mpg.de

DL: dirk.linke@tuebingen.mpg.de

ABSTRACT

Background

In Gram-negative bacteria, the outer membrane is composed of an asymmetric lipid bilayer of phopspholipids and lipopolysaccharides, and the transmembrane proteins that reside in this membrane are almost exclusively β-barrel proteins. These proteins are inserted into the membrane by a highly conserved and essential machinery, the BAM complex. It recognizes its substrates, unfolded outer membrane proteins (OMPs), through a C-terminal motif that has been speculated to be species-specific, based on theoretical and experimental results from only two species, *Escherichia coli* and *Neisseria meningitidis*, where it was shown on the basis of individual sequences and motifs that OMPs from the one cannot easily be over expressed in the other, unless the C-terminal motif was adapted. In order to determine whether this species specificity is a general phenomenon, we undertook a large-scale bioinformatics study on all predicted OMPs from 437 fully sequenced proteobacterial strains.

Results

We were able to verify the incompatibility reported between *Escherichia coli* and *Neisseria meningitidis*, using clustering techniques based on the pairwise Hellinger distance between sequence spaces for the C-terminal motifs of individual organisms. We noticed that the amino acid position reported to be responsible for this incompatibility between *Escherichia coli* and *Neisseria meningitidis* does not play a major role for determining species specificity of OMP recognition by the BAM complex. Instead, we found that the signal is more diffuse, and that for most organism pairs, the difference between the signals is hard to detect. Notable exceptions are the *Neisseriales*, and *Helicobacter spp*. For both of these organism groups, we describe the specific sequence requirements that are at the basis of the observed difference.

Conclusions

Based on the finding that the differences between the recognition motifs of almost all organisms are small, we assume that heterologous overexpression of almost all OMPs should be feasible in *E. coli* and other Gram-negative model organisms. This is relevant especially for biotechnolgoy applications, where recombinant OMPs are used e.g. for the development

of vaccines. For the species in which the motif is significantly different, we identify the residues mainly responsible for this difference that can now be changed in heterologous expression experiments to yield functional proteins.

KEYWORDS

Outer membrane β-barrel protein biogenesis, clustering, Hellinger distance, CLANS, species specificity, short linear motifs, GLAM2, C-terminal β-strand, BamA, β-barrel assembly machinery, Gram-negative bacteria, Outer membrane, Principal component analysis, Frequency plots

BACKGROUND

In Gram-negative bacteria, the cytoplasm is surrounded by inner membrane (IM) and outer membrane (OM), which are separated by an inter-membrane space, called the periplasm. Most of the newly synthesized proteome remains in the cytoplasm, but in addition, different machineries are involved in the translocation of non-cytoplasmic proteins to different subcellular localizations, including the inner or outer membrane, the periplasmic space, or the extracellular space. Some of these machineries recognize their substrate proteins by an N-terminal signal peptide (SP) for the translocation process, while other machineries are SP-independent. The IM, which is a phospholipid lipid bilayer, is mostly occupied by transmembrane α-helical proteins, by inner membrane lipoproteins on its periplasmic side, and by other membrane associated proteins on both sides of the membrane. In contrast, the asymmetric OM, which consists of phospholipids only in the inner leaflet of the membrane and lipopolysaccharides in the outer leaflet, is mostly occupied by transmembrane (outer membrane) β-barrel proteins, and by outer membrane lipoproteins on its periplasmic side [1].

The biogenesis of an outer membrane β-barrel protein (OMP) begins with the translocation of the newly synthesized, unfolded protein across the IM into the periplasm via the Sec translocation machinery, which requires a cleavable general SP. Once the unfolded OMP reaches the periplasm, it uses the SurA or Skp-DegP pathway to reach the OM. SurA, Skp and DegP are periplasmic chaperones, which interact with unfolded OMPs by protecting them from aggregation and thus help them to reach the OM [2-3]. It has been shown that the SurA pathway and the Skp/DegP pathway can work in parallel, but that the SurA pathway

plays an important role when the cell is under normal growth conditions, while under stress conditions, the Skp-DegP pathway plays the major role [4-5].

Once periplasmic chaperones deliver the OMPs to the OM, the folding and insertion of the protein into the membrane is mediated by the β-barrel assembly machinery (BAM), without an external energy source [6] such as ATP or ion gradients. This machinery involves an essential multi-domain protein, BamA (Omp85), which consists of a 16-stranded transmembrane β-barrel domain, and of a large periplasmic part that consists of five POTRA (polypeptide transport-associated) domains. BamA is highly conserved in Gram-negative bacteria and also has homologues in mitochondria (Sam50) and chloroplasts (Toc75-V) [2]. In addition, the BAM complex, at least in *E. coli*, consists of four lipoproteins, BamB, BamC, BamD and BamE, among which only BamD is essential and conserved in most Gram-negative bacteria [2]. Recent HMM-based sequence analysis by Anwari *et al* [7] showed that BamB and BamE are mainly present in α-, β- and γ-proteobacteria, while BamC is present only in β- and γ-proteobacteria. They also found a new lipoprotein subunit in the BAM complex, named BamF, which is present exclusively in α-proteobacteria.

The BAM complex recognizes OMPs as its substrates via binding to an amphipathic C-terminal β-strand of the unfolded β-barrel [8], but the exact binding mode is still not clear. It was suggested that C-terminal β-strand binds to BamD [9], once the unfolded OMPs are delivered to the BAM complex by periplasmic chaperones. But a recent BamC and BamD subcomplex crystal structure shows that the unstructured N-terminus of BamC binds to the proposed substrate binding site of BamD [4]. The C-terminal β-strand of an OMP β-barrel domain typically contains an aromatic residue at its C-terminus. It has been reported that deletion or substitution of this C-terminal residue negatively affects the biogenesis of OMPs [10-11]. Also, *in vitro* studies showed that the *E. coli* OM porin PhoE, when lacking its C-terminal Phe residue, fails to open the Omp85/BamA channel [8]. In both studies, overexpression of the mutant OMP was lethal to the cells. At lower concentration, the mutant protein was tolerated and got inserted into the membrane. This leads to the suggestion that a weak insertion signal other than the C-terminal residue or β-strand is present [8].

Robert *et al* [8] observed that the *N. meningitidis* OM porin PorA or its C-terminal β-strand did not open the *E. coli* Omp85/BamA channel, and the comparison of the C-terminal β-strands from *N. meningitidis* and *E. coli* OMPs showed a high preference of positive amino acids at the penultimate (+2) position in neisserial OMPs. When they mutated *E. coli* PhoE or its C-terminal β-strand, changing Gln for Lys at the +2 position, it did not open the channel any more; in contrast, a *Neisseria* PorA peptide with Gln instead of Lys increased the channel

activity considerably. These studies and the fact that high concentrations of neisserial OMPs were lethal in *E. coli* cells, lead to the conclusion that the C-terminal insertion signal is species-specific and that the residues at the +2 position were important for this phenomenon. The number of peptides/proteins used in the comparison in the study [8] was very low, compared to the total number of OMPs present in the *E. coli* or *N. meningitidis* genomes; moreover, the phenomenon was only compared between two organisms, one β- and one γ-proteobacterial species. Since neisserial OMPs could be expressed in *E. coli* at low expression rates, either the neisserial C-terminal insertion signal is weakly recognized by *E. coli* BAM complex, or other β-strands in the full length protein might act as a weak insertion signal.

Thus, there seems to be at least some overlap in the peptide recognition. The intention of this study was to use computational methods to quantify this overlap, and to find out whether the observed (partial) species specificity of the insertion signal is exhibited by all Gram-negative bacterial organisms.

RESULTS AND DISCUSSION

We identified 22,447 OMPs from 437 Gram-negative bacteria using PSORTb [12], CELLO [13] and HHomp [14] as described in the methods section. These OMPs can be classified into different outer membrane protein (OMP) classes/families based on their function and the number of β-strands present in them, as these two features are usually coupled [14-17]. We used HHomp [14] to classify the proteins into different OMP families. A brief summary of the OMP classification obtained from HHomp [14] for our data set is shown in Table 1. We then used ProfTMB [18] and PSIPRED [19] annotations to identify and extract the C-terminal β-strands from the OMPs. To evaluate the phenomenon of species specificity, we initially tried to cluster the C-terminal β-strands using different methods, such as sequence based clustering in CLANS [20] and organism-specific PSSM profile-based hierarchical clustering. Since the sequences were highly similar and very short, the results obtained from these methods were not helpful to our analysis. We then used chemical descriptors and represented each amino acid in the peptides by five-dimensional vectors, thus representing each 10-residue peptide as a 50-dimensional vector. Next, we used dimensionality reduction techniques (principal component analysis) to reduce the dimensions to 12 (the lowest number of dimensions that still contains most of the difference information, see Methods). We then used all peptide vectors from an organism to derive a multivariate Gaussian distribution, which we describe as the 'peptide sequence space' of the organism. The overlap between these multidimensional peptide sequence spaces (multivariate Gaussian distributions) was calculated using a statistical theory method, the Hellinger distance. As described in the methods section, the pairwise overlaps between organism sequence spaces were used to cluster them in CLANS [20].

Clustering of organisms based on C-terminal β-strands:

The pairwise comparison of the overlap between sequence spaces should help us to predict the similarity between the C-terminal insertion signal peptides, and how high the probability is that the protein of one organism can be recognized by the insertion machinery of another organism. When there is a complete overlap of sequence space between two organisms, we assume that all C-terminal insertion signals from one organism will be recognized and functionally expressed by another organism's BAM complex and vice-versa. When there is only little overlap between the sequence spaces of two organisms, we assume that only a small number of C-terminal insertion signals from one organism will be

recognized by another organism's BAM complex. When there is no overlap, we assume that there is a general incompatibility.

As described in the methods section, we examined the overlap of peptide sequence spaces between 437 Gram-negative bacterial organisms and used the pairwise overlap measurement to cluster the organisms. Since the C-terminal β-strands are highly conserved between all OMPs [21], it was very difficult to select a particular cut-off for the distance measure. Thus, the clustering was carried out using all the distance measures obtained from the calculations. In the resulting 2D cluster map (figure 1A), each node is one out of the 437 organisms, and they are colored based on the taxonomic classes (see the figure legend). During clustering with default clustering parameters in CLANS [20], the organisms tended to collapse into a single point, which illustrates that there is large overlap between the peptide sequence spaces. Thus, we introduced very high repulsion values and minimum attraction values in CLANS [20] during clustering. With these settings the organisms formed a central big cluster, but separated crudely according to their taxonomic classes. We repeated the clustering multiple times to ensure that this separation is reproducible. In the cluster map (figure 1A), β- and γ-Proteobacteria form two sub-clusters, separated by the α-Proteobacteria. The very few δ-Proteobacteria in our data set cluster in the periphery of the γ-proteobacterial cluster. In the cluster map, *E. coli* strains cluster along with other γ-Proteobacteria. Even though *Neisseria* species cluster along with other β-Proteobacteria, they form a sub-cluster and are found in the periphery of the β-proteobacterial cluster. Note also that in this map, *Helicobacter* species form a distinct cluster well separated from the rest of the organisms. This core cluster includes *H. pylori* strains, *H. acinonychis* and *H. felis,* but not *H. hepaticus* and *H. mustelae* species. The remaining ε-proteobacteria species are scattered in the periphery of the cluster map. The distinct cluster formed by most *Helicobacter* species demonstrates that the sequence spaces of *Helicobacter* species are significantly different from rest of the organisms. The neisserial cluster had only very few strong connections even with other β-proteobacterial organisms, which means the overlap or similarity of peptide sequence space between *Neisseriales* with rest of the β-Proteobacteria is comparatively low. When we used stringent thresholds for the distance measure, we noticed that the *Neisseria* and *Helicobacter* clusters started to move even further away from the center of the cluster map.

Control experiments for clustering: randomly shuffled peptide sequences lose the signal for clustering.

We noticed that the organisms seen at the periphery of the cluster map had a lower

overall number of peptides, while organisms with more peptides are typically seen at the center of the circle. The cluster map in figure 1B is colored based on the number of extracted peptides per organism. In figure 1B, there are 99 organisms which have ≤ 30 peptides (colored in pink), 77 organisms with 31 to 40 peptides (colored in blue), 136 organisms with 41 to 60 peptides (colored in green), 66 organisms with 61 to 80 peptides (colored in red), and 59 organisms with more than 80 peptides (colored in brown). Even though *H. pylori* strains have a comparably high number of peptides (43 to 51 peptides), they still form a separate cluster in the periphery of the cluster map; therefore there must be an underlying organism-specific signal from the contributing peptides at least in this case.

To confirm the presence of the organism-specific signal, we took peptides from all the organisms and shuffled the positions of their amino acids randomly, and derived a new similarity matrix as mentioned in the method section which we clustered in CLANS [20]. Figure 2A shows the results from this test, where one can notice the taxonomic specific separations were completely lost. The cluster map in figure 2B, colored based on the abundance of OMPs in an organism, shows that organisms with more peptides are in the center, and organisms with fewer peptides move to the outer rim of the cluster map. This test confirms that the there is a species-specific signal for which the position of the individual amino acids is important; this is lost when the residues in the peptides are shuffled randomly.

High preference of positively charged residues at the +2 position in *Neisseria* species:

The comparison of the C-terminal peptide sequences in the β-barrel of selected OMPs of *E. coli* and *N. meningitidis* peptides by Robert *el al* [8] showed a strong preference for positively charged amino acids (Arg and Lys) at the +2 position in neisserial OMPs, which led to the suggestion of a distinct species specificity of the C-terminal β-strand recognition. Since the comparison was made from 11 and 9 OMPs from *E. coli* and *N.meningitidis,* respectively, we wanted to confirm this with a larger set of OMPs from the same bacterial species. The frequency plots in figure 3A and 3B were created from 171 (*E. coli*) and 50 (*N.meningitidis*) unique C-terminal β-strands. Comparison between these plots demonstrates the high preference of Arg and Lys at the +2 position in neisserial OMPs. When we checked the frequency of amino acids at the +2 position for 22,447 peptides from all 437 organisms, we noticed that in the complete dataset, Arg and Lys are the top two preferred residues at the +2 position, and that they are present in 31.62% (3996 + 3102) of the peptides. A similar frequency of Arg and Lys (31.32% (2262 + 1794 out of 12,949 unique peptides)) is observed

when only taking unique peptides into account (i.e. when duplicates are removed from the database). Figure 4 shows the percentage of Arg and Lys at the +2 position in 437 organisms; in this plot, *Neisseria* strains stand apart even from other β-proteobacterial organisms, and also from all other proteobacterial organisms. *Neisseria* strains (and a few α-proteobacterial organisms) have more than 60% of peptides with positively charged residues at the +2 position. Note, though, that also in all other organisms, positive charges are abundant there; for example, different *Escherichia* strains also have 25-40% of peptides with Arg and Lys at the +2 position. Thus, when these proteins are expressed, the *Escherichia* BAM complex should be able to recognize proteins with positively charged residues at +2 positions. As a matter of fact, there is experimental evidence for the functional expression of OMPs with positively charged residues at the +2 position in *E. coli* [22].

High preference of Histidine at the +3 position in porins (16-stranded OMPs) from β-proteobacteria:

In the frequency plots (figure 5) generated for each taxonomic class of Proteobacteria, we observed that the frequency of amino acids in the +2 positions were comparable, with the possible exception of the *Neisseriae*. In contrast to that, we observed a prevalence (up to 57% frequency) of His at the +3 position for β-proteobacteria, while the other taxonomic classes shared a similar, low (<15%) frequency of His in that position (Figure 6). 80% of the peptides with His at the +3 position belong to the β-proteobacteria and more than 92% of these peptides stem from 16-stranded β-barrel proteins (Porins, denoted as the OMP.16 class by HHOmp). None of the *Escherichia* C-terminal β-strands in our database have His at the +3 position, and experiments by Robert *et al.* were done with a *Neisseria* PorA peptide with a His at the +3 position. This might be the true reason why *E. coli* BamA didn't recognize neisserial peptides. When we further examined the available structures of porins from *Neisseria*, and we found the His at the +3 position to be present in the trimerization interface of the porins. Since the vast majority of the His residues at the +3 position of the C-terminal motifs were from 16-stranded porins that typically trimerize, this position might be relevant for trimerization in neisserial porins.

High preference of Tyrosine at the +5 position in *Helicobacter* species:

The separate cluster formed by *Helicobacter* species was an interesting observation for us, because it forms a more distinct cluster than *Neisseria*. This means that the peptide sequence space of *Helicobacter* species is more different from the rest of the organisms than

even the one of *Neisseriales*. But the frequency plots (figure 7A and 7B), generated from unique peptides of all *Helicobacter* species and *H. pylori* strains respectively, did not show a strong preference for any amino acid at either the +2 position and the strong preference of Tyr at +3 position is common among the c-terminal insertion signals. But, we noticed an uncommon strong preference of Tyr at the +5 position. The presence of a hydrophobic residue is common at +5 positions, but the presence of aromatic hydrophobic amino acids (especially Tyr) at the +5 position are highly preferred in *H. pylori* strains compared to other organisms (figure 8A and 8B). Since the peptide sequence space depends upon the entire sequence, we cannot confirm that the separate cluster formed by the *H. pylori* is exclusively due to the residues at this one particular position. There is experimental evidence that the expression of various *H. pylori* OMPs in *E. coli* is problematic [23]. Fisher *et al.* noticed that as long as the expressed *H. pylori* OMP remains in the cytoplasm of *E. coli*, it is not lethal, but that once it is secreted to the periplasm by the Sec machinery, it becomes lethal to *E. coli*. They also mentioned - without showing data - that removal of the C-terminal β-barrel region resulted in toleration of the proteins in the periplasmic space. This probably means that the *E. coli* BAM complex didn't recognize the C-terminal β-strands of the *H. pylori* OMPs, and the subsequent aggregation of the OMPs in the periplasm and the blockage of the BAM complex lead to the lethality. The authors concluded that the difference in OM lipid composition of *Helicobacter*, which contains cholesteryl glycosides [24], might have imposed some structural constraints on the OMP structure, and that this structural change is not tolerated by other organisms resulting in the observed lethality of such constructs.

OMP class-specific and taxonomy class-specific signals:

We noticed that in some organisms, certain OMP classes of proteins are over-represented (see figure in additional file 2). Examples are the prevalence of 16-stranded β-barrels in the genomes of some β-proteobacteria and 22-stranded β-barrels in the genomes of some α-proteobacteria (see supplementary material S1). Moreover, of the 22,447 sequences in the data set, 33.82% (7591) sequences were annotated as OMP.nn by HHomp [14], which means there was no closely related homolog of known structure found for these proteins and thus, the number of β-strands in them is unknown. Thus, it is not possible to filter the dataset based on OMP class alone. But, as a control, we removed one OMP class at a time from the dataset and checked for differences in the clustering. When removing OMP.8 (figure 9A) and OMP.12 (figure 9B), two OMP classes that are not overrepresented in any of the taxonomy classes; this did not visibly affect the clustering. But when we removed the OMP.16 (figure

9C) or the OMP.22 (figure 9D) class, which have a high prevalence in β-proteobacteria and α-proteobacteria, respectively, this changed the clustering behavior of the respective taxonomic classes significantly; the organisms got scattered away from their position in the cluster compared to the situation in figure 1A. This shows that the over-representation of certain OMP classes can influence the peptide sequence space, but since the proteins from over-represented OMP classes still contribute to the real sequence space of the organisms, we decided not to correct for this effect and used all peptides from the organisms in our experiments.

We also examined whether there is a more general signal from OMP classes, other than the signal from the over-representation of an individual OMP class that would influence the observed organism-specific signal. For this, we separated the peptides from an organism based on the OMP classification and selected the entities which had more than five unique peptides for further analysis. From this, we created two data sets of entities; one data set containing organisms from all taxonomic classes, but with C-terminal insertion signals only from 22-stranded OMPs, and a second data set containing organisms only from γ-proteobacteria, but in which individual organisms were split into multiple entities, each representing an OMP class that contained more than five unique C-terminal insertion signals. We clustered these data sets separately and the resulting cluster maps are shown in figure 10A and 10B. In the cluster map in figure 10A, each node is an organism, but only the C-terminal insertion signals from 22-stranded OMP class were considered for the clustering. In this cluster map, all the organisms clustered based on their taxonomic classes. In the cluster map in figure 10B, all organisms are from γ-proteobacteria, but organisms with multiple OMP classes with more than five unique C-terminal insertion signals per class will result in multiple representative nodes. These nodes which belong to different OMP classes clustered based on the OMP classes. This confirms that there are independent contributions to the overall signal, from both the OMP classes and from taxonomy. Within one OMP class, there still is divergence in accordance with different taxonomic classes; but overrepresentation of a single OMP class in an organism influences the average motif of an organism.

CONCLUSION

In our study, we were able to reproduce the difference between *E. coli* and *Neisseria* C-terminal β-strands as found by Robert *et al.*, which suggests a species-specific insertion signal for OMPs. But in contrast to the earlier report, we show that positively charged amino acids at the +2 position can not be the reason for the experimentally observed species specificity between these organisms, as *Escherichia* also contains C-terminal β-strands with positively charged amino acids at the +2 position. Moreover, there is experimental evidence which shows the functional expression of a heterologous OMP, YadA of *Yersinia enterocolitica,* with a positively charged amino acids at the +2 position, in *E. coli* [22]. The neisserial PorA protein and the neisserial C-terminal β-strands used by Robert *et al.* contained His at the +3 position, which is common for many OMP.16 proteins from β-proteobacteria and is not found in *Escherichia* OMPs; this might be the true difference in the recognition of C-terminal β-strands by the *Escherichia* BAM complex. Furthermore we found that *Helicobacter* strains form a distinct cluster in the cluster map, which is due to their very different composition of C-terminal β-strands. There is experimental evidence showing that expression of *H. pylori* OMPs in *E. coli* is lethal, and that this lethality can be suppressed by removing the C-terminal strand. When we looked at the frequency motifs from *Helicobacter* strains we did not notice a strong preference of any amino acid at the +2 or the +3 position, however we observed a strong preference of Tyr at the +5 position, which is not common in *Escherichia* or other Proteobacteria. We assume that this position may play an important role in the rejection of these C-terminal β-strands by the *E. coli* BAM complex. The examples of *Neisseria* and *Helicobacter* show that different positions in the C-terminal recognition motif can be relevant for heterologous expression of OMPs. We predict that in certain group of species the highly preferred residues in certain positions of the C-terminal insertion signals are responsible for the inadequate recognition of the C-terminal insertion signals by the *E. coli* BAM complex. In the future, mutation studies will have to be performed to prove the importance of these residues in the recognition step in the OMPs biogenesis.

As a result of our study, we have shown that there is a large overlap between the signals from C-terminal insertion peptides of different organisms, which suggests that in most cases, heterologous expression should be possible. OMPs can fold *in vitro* even without the help of any other proteins [25]. The BAM complex is an enzyme that makes the folding of OMPs into the outer membrane more efficient by increasing the reaction rate of a natural process. Enzymes modify reaction rates by changing the reaction route to lower the activation energy, and binding/recognition is part of this changed route. Thus, it is also important to

consider expression rates: poor recognition might still lead to properly folded OMPs in the outer membrane of a heterologous host at low expression rates. But under overexpression conditions, the BAM machinery can probably not cope with poorly recognized signals that would lead to lower overall folding rates (considering that recognition is the first and probably in some cases rate-limiting step of the folding process). Different classes of OMPs have different folding rates, where small OMPs fold faster and more efficiently (again *in vitro*) than larger ones, which might explain why large OMPs seem to depend more heavily on an intact BAM machinery than small ones [26-27].

Since there are two different signals that contribute to the observed average motifs, from OMP class and from taxonomy, it is problematic to use averaged motifs or sequence logos to determine the compatibility of a given protein-organism pair. The main problem here is the overrepresentation of certain OMP classes in some organism groups; this overrepresentation shifts the average signals. It is more useful to determine for an individual C-terminal motif form a protein to be expressed, whether it is also present in any of the OMPs of the host organism.

The taxonomy-based specificity we observed here based on sequence space depends upon the entire peptide sequence, but at the functional level, these peptides are recognized based on the interacting residue positions in the C-terminal insertion signal peptide. The PDZ domain of the bacterial periplasmic stress sensor, DegS, also recognizes the C-terminal YxF motif in the last β strand of misfolded OMPs. This leads to the activation of the proteolytic pathway and the expression of DegP, which degrades misfolded OMPs [28-29]. Since the C-terminal β-strand is recognized by both the PDZ domain of the DegS protein and by the BAM complex, studying the co-evolution of interacting residues in both cases would help in understanding the divergence of the C-terminal β-strands between different Gram-negative bacterial organisms. Unfortunately, co-crystal structures of the BAM complex with its substrates are not available yet. With more experimental evidence about the substrate recognition sites for the C-terminal insertion signal peptide in the BAM complex, the co-evolution of the interacting amino acids can hopefully be studied in the future, which may shed more light on into the evolution of the BAM machinery in different Proteobacteria, and on its ability to recognize heterologous substrates for biotechnology applications.

METHODS

Predicting Outer membrane β-barrel proteins:

In a previous study [30] to annotate the subcellular localizations (SCLs) for the proteomes of 607 Gram-negative bacteria, we developed the program/database ClubSub-P, in which we used programs like CELLO [13], PSORTb [12] and HHomp [14] to annotate OMPs. CELLO [13] and PSORTb [12] use support vector classifiers to annotate different SCLs of query sequences and are much faster than HHomp [14] which uses HMM-HMM-based search algorithms to predict and classify OMPs. Thus we used CELLO and PSORTb to scan all the sequences in the clusters of the ClubSub-P database. A random protein was selected from a cluster where CELLO or PSORTb had a positive hit for an outer membrane protein, and the sequence was analyzed with HHomp. When HHomp predicted a protein with more than 90% probability to be an OMP, we considered all the proteins in the cluster to be OMPs. We in addition selected all singleton sequences with positive prediction from CELLO or PSORTb and analyzed them with HHomp.

Finding the C-terminal β-strands:

HHomp annotates/classifies OMPs based on the number of β-stands present in them. HHomp calculates/predicts this from homologous structures of OMPs. We transferred this annotation from the best hit in HHomp runs to the query sequences. HHomp also annotates secondary structure and β-barrel strand predictions using PSIPRED [19] and ProfTMB [18], which was used to extract the C-terminal (last) β-strand/motif for each OMP. The last β-strand predicted by ProfTMB [18] was extracted as the C-terminal motif from representative sequences and singletons, and further filters were applied to reduce the false positive rate; 1) 70% of the amino acids in the motif should have a β-strand prediction from PSIPRED [19], 2) If the C-terminal of the protein is more than 4 residues away from the C-terminus of the motif, we extended the predicted motif by up to 4 amino acids to find an aromatic hydrophobic residue [F,Y,W], else we extended the C-terminus of the motif to the end of the protein itself. 3) Additionally, if the motif length was less than 10 residues, we extended the motif towards its N-terminus. 4) Furthermore with the regular expression [^C][YFWKLHVITMADGRE][^C][YFWKLHVITMADGRE][^C][YFWKLHVITMADGRE][^C].[^C][YFWHILM] (an updated version of BOMP[31] C-terminal pattern), we searched for the existence of the alternating hydrophobic pattern in the motif which is typical for transmembrane β-strands.

Using the information from this representative C-terminal motif, we extracted C-terminal motifs from the rest of the sequences in the clusters. We used MAFFT [32] to align the sequences from the cluster, and used the start and end coordinates of the C-terminal motif discovered above in the representative sequences randomly selected from the clusters. Motifs were extended on the both sides, in cases where we encountered gaps in the alignment. The gaps were removed and then resulting motifs were subjected to alternating hydrophobic pattern matching.

The peptides we collected vary in length from 10 to 21 residues (only six of the peptides were longer than 21). We then applied GLAM2 [33], a gapped motif discovery algorithm, to find the strongest motif with a length of 10 from this dataset. We found 24,626 motif instances in 25,454 sequences, and only 232 motifs in this alignment had gaps. The gapped motifs were removed before further analysis. 20,135 of the motif instances were C-terminal to the protein itself (which means there were no additional domains at the C-terminal end of the β-barrel proteins). 437 organisms had more than 20 unique C-terminal β-strands, ranging from 21 to 171 peptides in different organisms. In total, the 437 organisms yielded 22,447 peptides, of which 12,949 are unique peptides.

Sequence based clustering:

Since all of the peptides are 10 amino acids in length by default, we used the PAM30 substitution matrix for an all-against-all BLAST, with an E-value cut-off of 1000 and used the pairwise P-values to cluster the sequences in CLANS [20].

PSSM profile-based hierarchical clustering:

The relative frequencies of the 20 amino acids were calculated for all 10 positions in the peptides from an organism. To obtain odds scores, the relative frequencies were simply divided by each residue's background frequency, which was calculated by shuffling the amino acid sequence in all the peptides from all organisms, and log base 2 was applied to obtain a PSSM matrix. The 20 x 10 PSSM matrices obtained for each organism were stored in a single 437x200 PSSM matrix, and correlation distances were calculated between each organism and agglomerative hierarchical clustering (average method) was performed via the pvclust [34], which calculates two types of *p*-values, AU (Approximately Unbiased) *p*-value and BP (Bootstrap Probability) value to indicate the likelihood of the cluster formation.

Peptide sequence space-based clustering:

Chemical descriptors:

To generate a peptide sequence space, each amino acid in the peptide sequences was represented by five chemical descriptors that are the first five principal components derived from 26 physiochemical descriptor variables using dimensionality reduction techniques [35]. The initial 26 physiochemical descriptor variables include the molecular weight, experimentally determined retention values from seven thin-layer chromatography runs, van der Waals volume of the side chain, three nuclear magnetic resonance shift variables, log *P*, six variables for semiempirical molecular orbitals, three variables for total, polar and nonpolar surface area, two variables for side chain charge and two variables for hydrogen bond donor and acceptor [35]. The five principal components derived from these 26 variables contain the maximal variations in the data set and they can be interpreted as the size, polarizability, and the lipophilic, steric, and electronic properties of all the amino acids [35]. The amino acid descriptors were originally derived for use as design variables in peptide design, and in the construction of combinatorial libraries to effectively search chemical property space [35]. Here we used them to describe the space occupied by the C-terminal β-strands and to measure how strongly peptide sequences of different organisms overlap. Using the chemical descriptors, each amino acid in the peptide was converted into a 5-dimensional vector; thereby, each 10aa peptide was represented as a 50-dimensional vector. Thus, the whole set of 22,447 peptides were converted to a 22,447 x 50 matrix.

Principal component analysis:

Since the dimensionality of the data set (50) is larger than the sample size (minimum 21 peptides per organism), the dimensionality of the peptide vectors had to be reduced below the sample size (i.e., below 21 in our dataset) for further statistical analysis [36]. Principal component analysis (PCA) is a mathematical technique to reduce the dimensionality of data sets, while retaining most of the variation in the data set. This is achieved by projecting the original data vectors along the directions of maximal variation, called principal components (PCs). The first PC captures the maximum variation; the variation associated with consecutive PCs decreases rapidly. Thus, the original data set can be mapped into a lower dimensional space by projecting the original data on those PCs representing most of the variation [36-37]. We used PCA to reduce the dimensionality of our peptide sequences (22,447 x 50 matrix) by projecting the 50 dimensional chemical descriptor vectors onto the first 12 principal components, which represent 69.05% of the total variation in the data. We

thereby obtained a 22,447 x 12 matrix that did not suffer from any problems in sample size.

Multivariate Gaussian fitting and Hellinger distance:

Next, we fit a multivariate Gaussian distribution for each individual organism by calculating a 12-dimensional mean vector and covariance matrix, (e.g., for *E. coli* 536 which has 66 unique peptides, the Gaussian will be fitted based on a 66 x 12 matrix).

The Euclidean distance between means of peptide sequence spaces is not suitable for measuring the similarity between the C-terminal β-strands of different organisms. Instead, the similarity measure should also represent how strongly their associated sequence spaces overlap. To achieve this we used the Hellinger distance between the fitted Gaussian distributions [38]. In statistical theory, the Hellinger distance measures the similarity between two probability distribution functions, by calculating the overlap between the distributions. For a better understanding, figure 11 illustrates the difference between the Euclidean distance and the Hellinger distance for one-dimensional Gaussian distributions. The Hellinger distance, $D_H(Org_1,Org_2)$, between two distributions Org1(x) and Org2(x) is symmetric and falls between 0 and 1. $D_H(Org_1, Org_2)$ is 0 when both distributions are identical; it is 1 if the distributions do not overlap [39]. Therefore we have for the squared Hellinger distance $D_H^2(Org_1, Org_2) = 1 - \text{overlap}(Org1,Org2)$. The following equation (1) was derived to calculate the pairwise Hellinger distance between the multivariate Gaussian distributions, Org_1 and Org_2, where μ_1 and μ_2 are the mean vectors and Σ_1 and Σ_2 are the covariance matrices of Org_1 and Org_2, and *d* is the dimension of the sequence space, i.e. *d*=12.

$$D_H(Org_1, Org_2) = \sqrt{1 - 2^{d/2} \left(\frac{\det(\Sigma_1)\ \det(\Sigma_2)}{\det(\Sigma_1 + \Sigma_2)^2} \right)^{1/4} \exp\left\{ -\frac{1}{4}(\mu_1 - \mu_2)^T (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2) \right\}}$$

(1)

CLANS:

Next, the Hellinger distance was used to define a dissimilarity matrix for all pairs of organisms. The dissimilarity matrix was converted to P-values, which were then used as input in CLANS [20] to compute a cluster map showing all organisms. CLANS is a graph-based clustering method that represents sequences as nodes. All nodes are connected by weighted edges where the pairwise similarity between the sequences determines the strength of the weight [20]. In our study, individual organisms were considered as nodes and the weight of

the edges connecting the nodes was based on the pairwise Hellinger distance (pairwise overlap of sequence space) between the organisms. Hence stronger connections represent a larger overlap/similarity between the peptide sequence spaces, while organisms with high divergence in their C-terminal motifs are only weakly connected or completely disconnected in the cluster map. Initially the nodes are randomly placed in a 2D space and experience attraction forces according to how strongly they are connected with the other nodes. In an iterative refinement scheme, nodes move towards similar nodes with an attractive force proportional to the similarity between them. A small, overall repulsive force is applied to all pairs of nodes to keep them from collapsing into a single node. Since CLANS [20] uses non-deterministic dynamics, each run performed with the same dataset will result in a similar but not necessarily identical clustering. Thus, multiple clustering runs were performed to check the reproducibility of the final clustering. Because initial tests showed that with the default attraction and repulsion values nodes (organisms) were collapsing, we used very small attraction values (up to 0.1) and high repulsion values (up to 500) to avoid collapse of nodes and to obtain visually better clusters.

Frequency plot:

The WebLogo [40] online tool was used to create the frequency plots, using custom colors. Only unique peptide sequences were used to generate all the frequency plots. The amino acid percentage plots were created using R version 2.13.1 [41].

AUTHOR'S CONTRIBUTIONS

NP generated and analyzed the data. MH provided the initial script for pairwise Hellinger distance calculation. DL conceived the initial idea about the project and helped in drafting the manuscript. NP wrote the manuscript, MH and DL read and improved the manuscript. All authors approved the manuscript.

ACKNOWLEDGEMENTS

We are grateful for helpful discussions with Vikram Alva, Iwan Grin, Jack Leo and other department members; continuing support by the Max Planck Society, and specifically by Andrei Lupas, is gratefully acknowledged.

COMPETING INTEREST

There is no competing interest.

REFERENCES

1. Silhavy TJ, Kahne D, Walker S: **The bacterial cell envelope.** *Cold Spring Harbor Perspectives in Biology* 2010, **2:**a000414.
2. Knowles TJ, Scott-Tucker A, Overduin M, Henderson IR: **Membrane protein architects: the role of the BAM complex in outer membrane protein assembly.** *Nature reviews Microbiology* 2009, **7:**206-214.
3. Bos MP, Robert V, Tommassen J: **Biogenesis of the gram-negative bacterial outer membrane.** *Annual review of microbiology* 2007, **61:**191-214.
4. Kim KH, Aulakh S, Paetzel M: **The bacterial outer membrane β-barrel assembly machinery.** *Protein Science* 2012,**21:**751-768
5. Sklar JG, Wu T, Kahne D, Silhavy TJ: **Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in Escherichia coli.** *Genes & Development* 2007, **21:**2473.
6. Hagan CL, Kim S, Kahne D: **Reconstitution of outer membrane protein assembly from purified components.** *Science (New York, NY)* 2010, **328:**890-892.
7. Anwari K, Webb CT, Poggio S, Perry AJ, Belousoff M, Celik N, Ramm G, Lovering A, Sockett RE, Smit J, Jacobs-Wagner C, Lithgow T: **The evolution of new lipoprotein subunits of the bacterial outer membrane BAM complex.** *Molecular microbiology* 2012, **84:**832-844.
8. Robert V, Volokhina EB, Senf F, Bos MP, Van Gelder P, Tommassen J: **Assembly factor Omp85 recognizes its outer membrane protein substrates by a species-specific C-terminal motif.** *PLoS biology* 2006, **4:**e377.
9. Sandoval CM, Baker SL, Jansen K, Metzner SI, Sousa MC: **Crystal Structure of BamD: An Essential Component of the β-Barrel Assembly Machinery of Gram-Negative Bacteria.** *Journal of molecular biology* 2011, **409:**348-357.
10. Struyvé M, Moons M, Tommassen J: **Carboxy-terminal phenylalanine is essential for the correct assembly of a bacterial outer membrane protein.** *Journal of Molecular Biology* 1991, **218:**141-148.
11. Hendrixson DR, De La Morena ML, Stathopoulos C, St Geme Iii JW: **Structural determinants of processing and secretion of the Haemophilus influenzae Hap protein.** *Molecular Microbiology* 1997, **26:**505-518.
12. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FSL: **PSORTb 3.0: improved protein subcellular localization**

prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 2010, **26:**1608-1615.

13. Yu C-S, Chen Y-C, Lu C-H, Hwang J-K: **Prediction of protein subcellular localization.** *Proteins* 2006, **64:**643-651.
14. Remmert M, Linke D, Lupas AN, Söding J: **HHomp--prediction and classification of outer membrane proteins.** *Nucleic acids research* 2009, **37:**W446-451.
15. Koebnik R, Locher KP, Van Gelder P: **Structure and function of bacterial outer membrane proteins: barrels in a nutshell.** *Molecular Microbiology* 2000, **37:**239-253.
16. Hritonenko V, Stathopoulos C: **Omptin proteins: an expanding family of outer membrane proteases in Gram-negative Enterobacteriaceae (Review).** *Molecular Membrane Biology* 2007, **24:**395-406.
17. van den Berg B, Black PN, Clemons WM, Rapoport TA: **Crystal Structure of the Long-Chain Fatty Acid Transporter FadL.** *Science* 2004, **304:**1506-1509.
18. Bigelow HR, Petrey DS, Liu J, Przybylski D, Rost B: **Predicting transmembrane beta-barrels in proteomes.** *Nucleic Acids Research* 2004, **32:**2566-2577.
19. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices.** *Journal of Molecular Biology* 1999, **292:**195-202.
20. Frickey T, Lupas A: **CLANS: a Java application for visualizing protein families based on pairwise similarity.** *Bioinformatics (Oxford, England)* 2004, **20:**3702-3704.
21. Remmert M, Biegert A, Linke D, Lupas AN, Söding J: **Evolution of outer membrane beta-barrels from an ancestral beta beta hairpin.** *Molecular biology and evolution* 2010, **27:**1348-1358.
22. Lehr U, Schütz M, Oberhettinger P, Ruiz-Perez F, Donald JW, Palmer T, Linke D, Henderson IR, Autenrieth IB: **C-terminal amino acid residues of the trimeric autotransporter adhesin YadA of Yersinia enterocolitica are decisive for its recognition and assembly by BamA.** *Molecular Microbiology* 2010, **78:**932-946.
23. Fischer W, Schwan D, Gerland E, Erlenfeld GE, Odenbreit S, Haas R: **A plasmid-based vector system for the cloning and expression of Helicobacter pylori genes encoding outer membrane proteins.** *Molecular and General Genetics MGG* 1999, **262:**501-507.
24. Hirai Y, Haque M, Yoshida T, Yokota K, Yasuda T, Oguma K: **Unique cholesteryl glucosides in Helicobacter pylori: composition and structural analysis.** *Journal of*

Bacteriology 1995, **177:**5327-5333.

25. Kleinschmidt J: **Membrane protein folding on the example of outer membrane protein A ofEscherichia coli.** *Cellular and Molecular Life Sciences* 2003, **60:**1547-1558.
26. Bos MP, Robert V, Tommassen J: **Functioning of outer membrane protein assembly factor Omp85 requires a single POTRA domain.** *EMBO reports* 2007, **8:**1149-1154.
27. Kim S, Malinverni JC, Sliz P, Silhavy TJ, Harrison SC, Kahne D: **Structure and function of an essential component of the outer membrane protein assembly machine.** *Science* 2007, **317:**961-964.
28. Walsh NP, Alba BM, Bose B, Gross CA, Sauer RT: **OMP Peptide Signals Initiate the Envelope-Stress Response by Activating DegS Protease via Relief of Inhibition Mediated by Its PDZ Domain.** *Cell* 2003, **113:**61-71.
29. Meltzer M, Hasenbein S, Mamant N, Merdanovic M, Poepsel S, Hauske P, Kaiser M, Huber R, Krojer T, Clausen T, Ehrmann M: **Structure, function and regulation of the conserved serine proteases DegP and DegS of Escherichia coli.** *Research in microbiology* 2009, **160:**660-666.
30. Paramasivam N, Linke D: **ClubSub-P: Cluster-based subcellular localization prediction for Gram-negative bacteria and Archaea.** *Frontiers in Microbiology* 2011, **2:**218.
31. Berven FS, Flikka K, Jensen HB, Eidhammer I: **BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria.** *Nucleic Acids Research* 2004, **32:**W394--W399.
32. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Briefings in Bioinformatics* 2008, **9:**286-298.
33. Frith MC, Saunders NFW, Kobe B, Bailey TL: **Discovering Sequence Motifs with Arbitrary Insertions and Deletions.** *PLoS Computational Biology* 2008, **4:**e1000071.
34. Suzuki R, Shimodaira H: **Pvclust: an R package for assessing the uncertainty in hierarchical clustering.** *Bioinformatics* 2006, **22:**1540-1542.
35. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S: **New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids.** *Journal of Medicinal Chemistry* 1998, **41:**2481-2491.

36. Ma S, Dai Y: **Principal component analysis based methods in bioinformatics studies.** *Briefings in bioinformatics* 2011, **12:**714-722.
37. Ringnér M: **What is principal component analysis?** *Nature Biotechnology* 2008, **26:**303-304.
38. Vajda I: **Theory of statistical inference and information.** *Kluwer Academic, Dodrecht, The Netherlands* 1989.
39. Shutin D, Zlobinskaya O: **Application of information-theoretic measures to quantitative analysis of immunofluorescent microscope imaging.** *Computer methods and programs in biomedicine* 2010, **97:**114-129.
40. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator.** *Genome Research* 2004, **14:**1188-1190.
41. Ihaka R, Gentleman R: **R: A language for data analysis and graphics.** *Journal of Computational and Graphical statistics* 1996**:**299-314.

FIGURE LEGENDS

Figure 1: Cluster map based on 437 sequenced Gram-negative organisms. In the cluster map each node represents one organism. The Hellinger distance was used to calculate the pairwise overlap between the multi-dimensional peptide sequence spaces of organisms. The calculated similarity or overlap was used to cluster the organism in CLANS. Figure 1A is colored by taxonomic class and figure 1B is colored by the number of peptides in each organism.

Figure 2: CLANS cluster map of randomly shuffled peptides from 437 organisms. Figure 2A is colored by taxonomic class and figure 2B is colored by the number of peptides in an organism. Colors are similar to figure 1.

Figure 3: Frequency plots derived from unique C-terminal insertion signal peptides for *Escherichia* (figure 3A) and *Neisseria* (figure 3B) strains. Frequency plots were made from 188 unique peptides of 31 *Escherichia* strains and 50 unique peptides of 7 *Neisseria* strains. The +2 position is indicated by the arrow in the figure. *Escherichia* strains (figure 3A) have no strong preference for any amino acid at the +2 position, whereas *Neisseria* strains (figure 3B) have a strong preference for positively charged amino acids (Arg and Lys) at the +2 position. Hydrophobic residues are colored in blue and polar residues are colored in red.

Figure 4: Percentage of Arg and Lys at +2 positions. We calculated the percentage of Arg and Lys residues at the +2 position from all unique peptides from the 437 organisms; color is based on taxonomic class. The *Neisseria* strains show a high preference for positively charged amino acids at the +2 position compared to other organisms.

Figure 5: Frequency plots of C-terminal β-strands from Proteobacteria. Frequency plots generated from unique peptides of α-proteobacteria are shown in figure 5A, of β-Proteobacteria in figure 5B, of γ-Proteobacteria in figure 5C, of δ-Proteobacteria in figure 5D and of ε-Proteobacteria in figure 5E. The frequency plots are overall very similar; an eception is the high frequency of His at the +3 position in β-Proteobacteria and of Tyr at the +5 position in ε-Proteobacteria.

Figure 6: Frequency of His at the +3 position. The percentage of His at +3 was calculated from all unique peptides from 437 organisms. A high preference for His at +3 is observed for 16-stranded OMPs of β-Proteobacteria. Since there is a high number of 16-stranded OMPs in *Burkholderia* strains (see additional file 1 and additional file 2), they were also annotated in the plot.

Figure 7: Frequency plot of unique C-terminal β-strands from *Helicobacter* species. 163 unique C-terminal insertion signals from 14 *Helicobacter* strains were used to generate this plot. The +5 position which has the strong preference of Tyr is marked with the arrow.

Figure 8: The percentage of Tyr (figure 8A) and aromatic hydrophobic amino acids (figure 8B) at the +5 position. For figure 8A, we calculated the percentage of Tyr at the +5 position from all unique peptides from 437 organisms amd for figure 8B, we calculated the frequency of Tyr, Phe and Trp at the +5 position from all unique peptides from 437 organisms. In both plots *Helicobacter* strains shows a high preference of Tyr and aromatic amino acids at the +5 position.

Figure 9: Control experiments to show the influence of overrepresented OMP classes. OMP classes OMP.8 (figure 9A), OMP.12 (figure 9B), OMP.16 (figure 9C) and OMP.22 (figure 9D) were removed and only organisms with more than 20 unique peptides were used in the clustering. Peptides belonging to OMP.nn and OMP.hypo (OMPs with unknown strand

number and function) were not removed from the data set during the control experiments. Color legends are similar to the figure 1A.

Figure 10: CLANS cluster map of OMP-Organism class based entities. In figure 10A and figure 10B, each node is a representative of OMP-Organism entities that have more than five unique peptides of a single OMP class from an individual organism. In figure 10A, entities are only from the OMP.22 class, which includes entities from all proteobacterial taxonomic classes. In figure 10B, entities are only from γ-Proteobacteria and include different OMP classes.

Figure 11: Illustration of the difference between the Euclidean distance and the Hellinger distance for one-dimensional Gaussian distributions. Two Gaussian distributions are shown as black lines for different choices of μ and σ. The grey area indicates the overlap between both distributions. $|\mu 1-\mu 2|$ is the Euclidean distance between the centers of the Gaussians, D_H is the Hellinger distance (equation 1). Both values are indicated in the title of panels A-D. A: For $\mu 1 = \mu 2 = 0$, $\sigma 1 = \sigma 2 = 1$, the Euclidean distance and the Hellinger distance are both zero. B: For $\mu 1 = \mu 2 = 0$, $\sigma 1 = 1$, $\sigma 2 = 5$ the Euclidean distance is zero, whereas the Hellinger distance is larger than zero because the distributions do not overlap perfectly (the second Gaussian is wider than the first). C: For $\mu 1 = 0$, $\mu 2 = 5$, $\sigma 1 = \sigma 2 = 1$, the Euclidean distance is five, whereas the Hellinger distance almost attains its maximum because the distributions only overlap little. D: For $\mu 1 = 0$, $\mu 2 = 5$, $\sigma 1 = 1$, $\sigma 2 = 5$, the Euclidean distance is still five as in C because the means did not change. However, the Hellinger distance is larger than in C because the second Gaussian is wider, which leads to a larger overlap between the distributions.

TABLES

Table 1: Dataset classified based on OMP class. The OMP class of a protein is determined by the number of β-strands present in the them, which was predicted by HHomp [14] based on the homologous relation to an OMP structure, when HHomp couldn't find a homologous structure, it classifies the proteins in OMP.nn. OMP.hypo proteins belong to the class of hypothetical proteins [14].

Table 1Dataset classified based on OMP class

OMP class	**# of β-strands**	**Total # of peptides**	**OMP class found in # of organisms in different proteobacteria class**					**Function/Protein family**
			α	**β**	**γ**	**δ**	**ε**	
OMP.8	8	2300	71	77	227	24	10	Membrane anchors [15]
OMP.10	10	95	5	2	66	2	2	Bacterial proteases [16]
OMP.12	12	1550	60	75	212	18	10	Integral membrane enzymes [15]
OMP.14	14	572	47	38	221	20	22	Long chain fatty acid transporter [17]
OMP.16	16	2477	41	86	210	23	8	General porins [15]
OMP.18	18	327	2	14	134	7	1	Substrate specific porins [15]
OMP.22	22	7462	71	86	231	25	23	TonB-dependent receptors [15]
OMP.nn	Not known	7591	71	86	231	26	23	-
OMP.hypo	Not known	73	2	18	33	9	1	-

ADDITIONAL FILES

Additional file 1

The table in the additional file 1 lists the number of OMPs in an organism present in different OMP classes.

Additional file 2

The figure in the additional file 2 shows the number the over representation of OMP.16 proteins among β-proteobacteria and OMP.22 among α-proteobacteria.

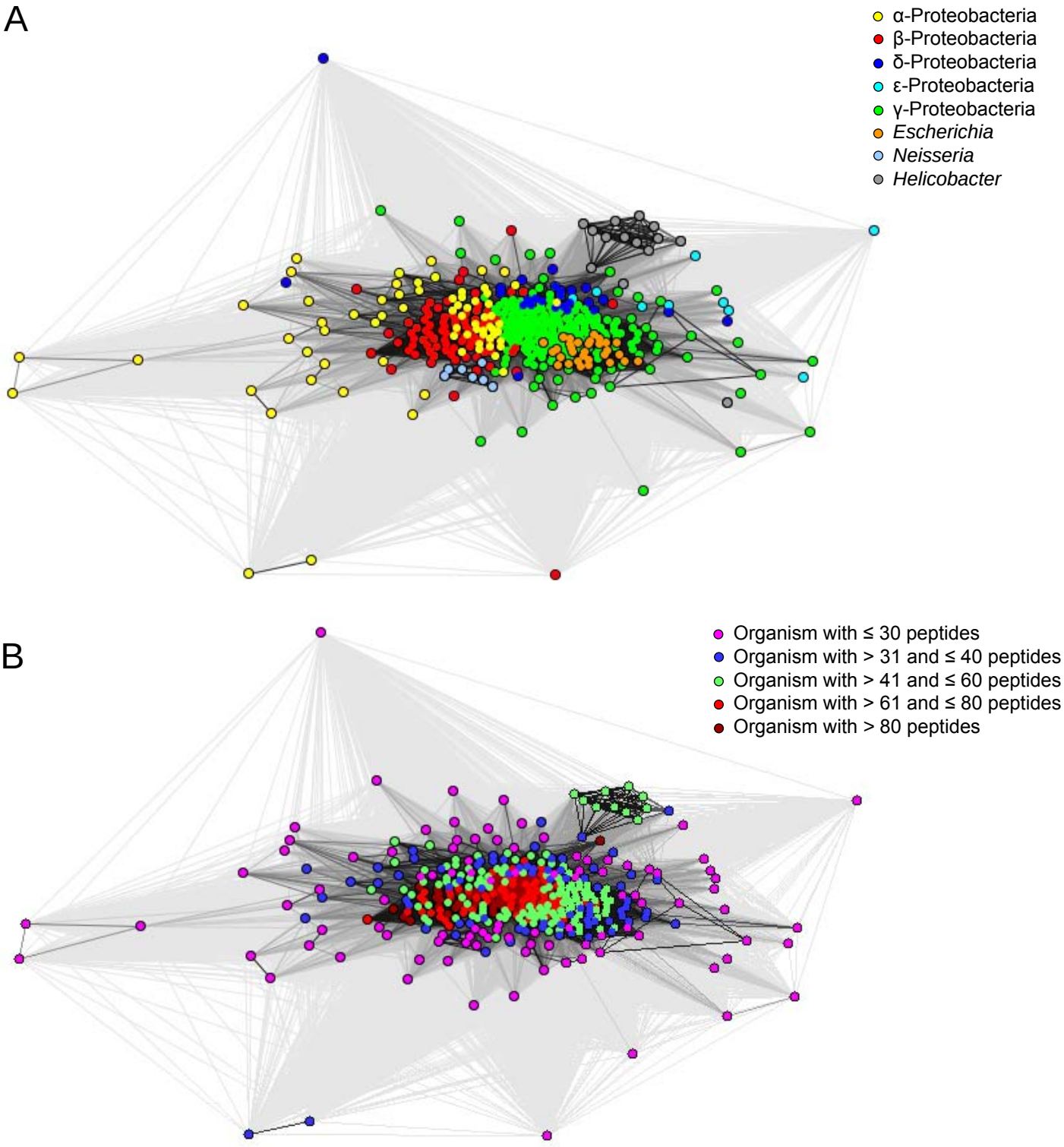

A
α-Proteobacteria
β-Proteobacteria
δ-Proteobacteria
ε-Proteobacteria
γ-Proteobacteria
Escherichia
Neisseria
Helicobacter
B
Organism with ≤ 30 peptides
Organism with > 31 and ≤ 40 peptides
Organism with > 41 and ≤ 60 peptides
Organism with > 61 and ≤ 80 peptides
Organism with > 80 peptides

Figure 1

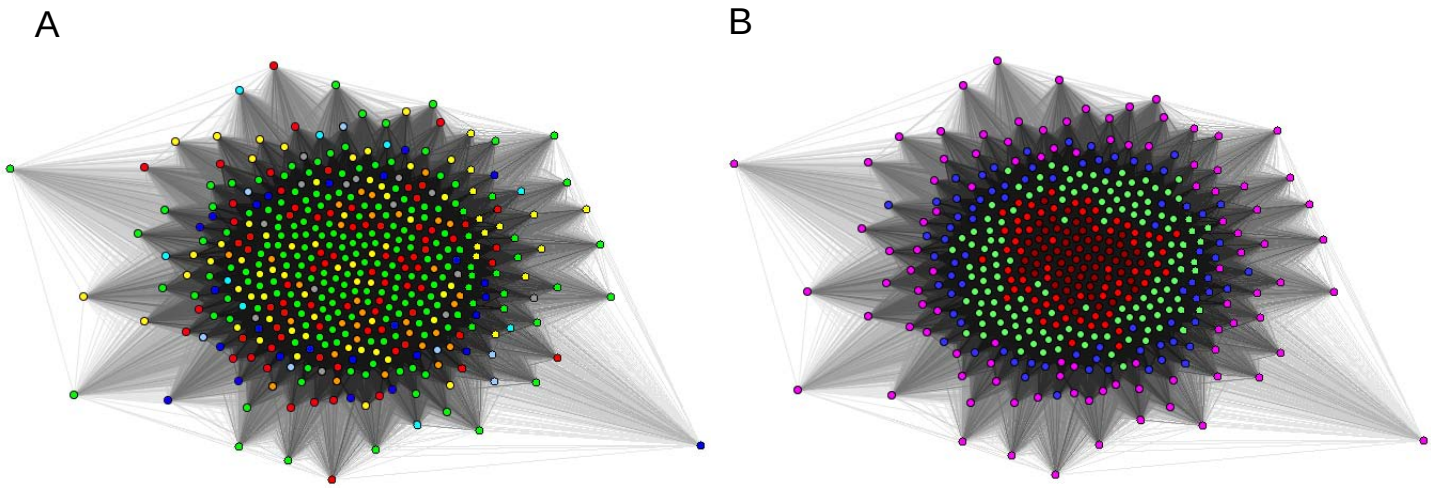


Figure 2

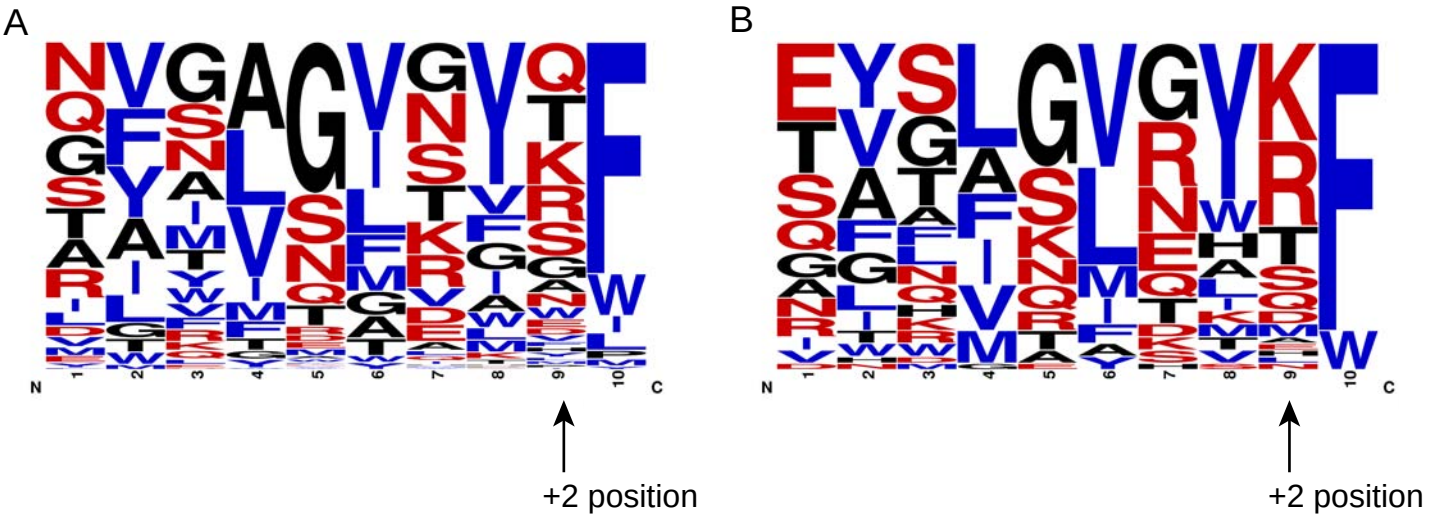


Figure 3

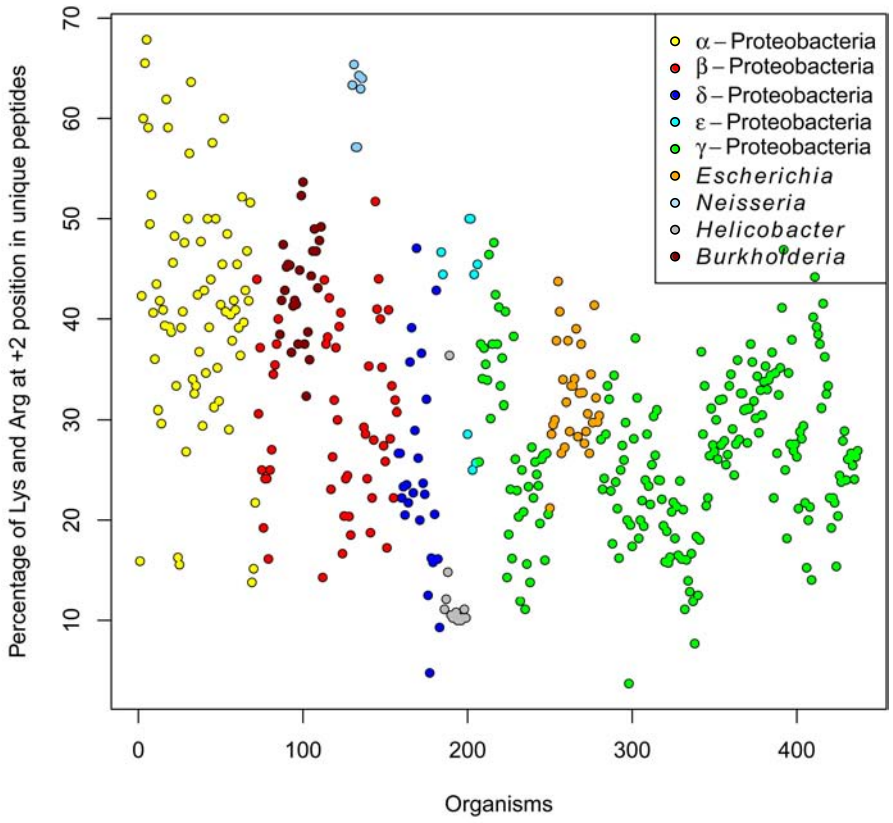
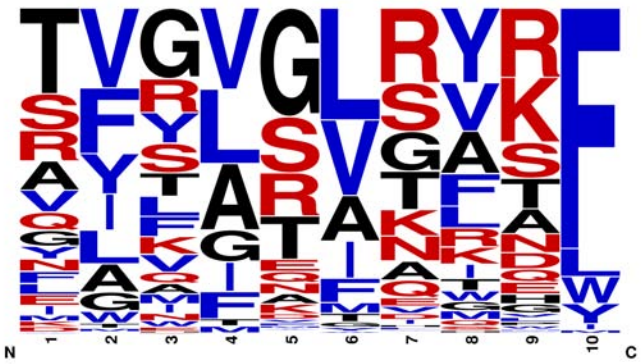
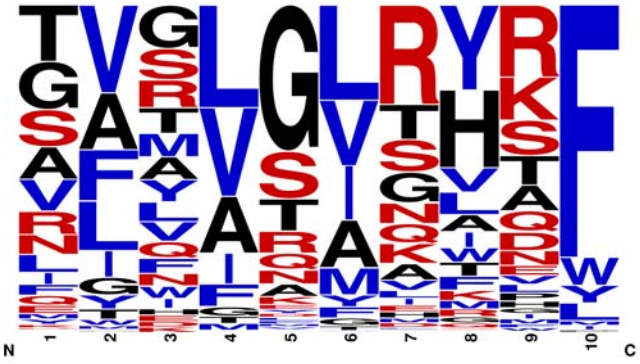

α – Proteobacteria
β – Proteobacteria
δ – Proteobacteria
ε – Proteobacteria
γ – Proteobacteria
Escherichia
Neisseria
Helicobacter
Burkholderia
Percentage of Lys and Arg at +2 position in unique peptides
10
20
30
40
50
60
70
0
100
200
300
400
Organisms

Figure 4

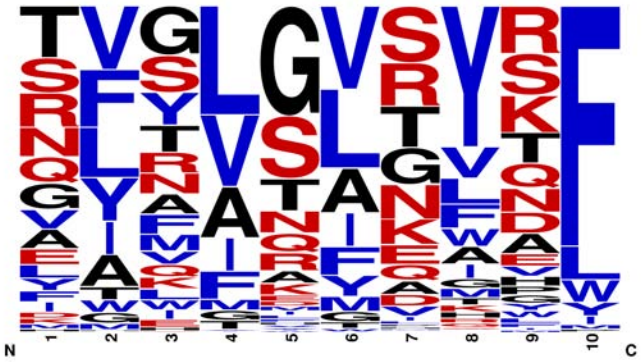
A



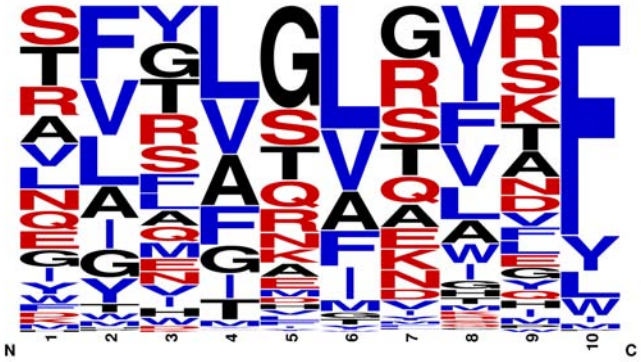
B



C



D



E

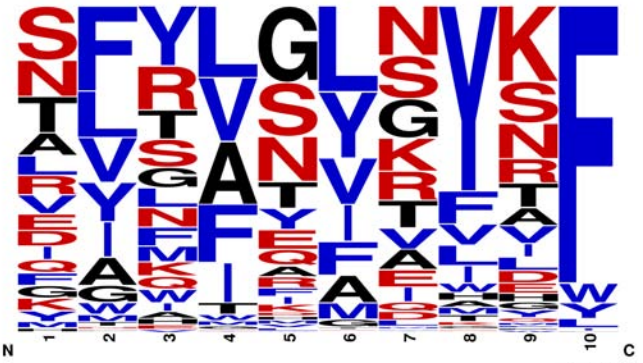


Figure 5

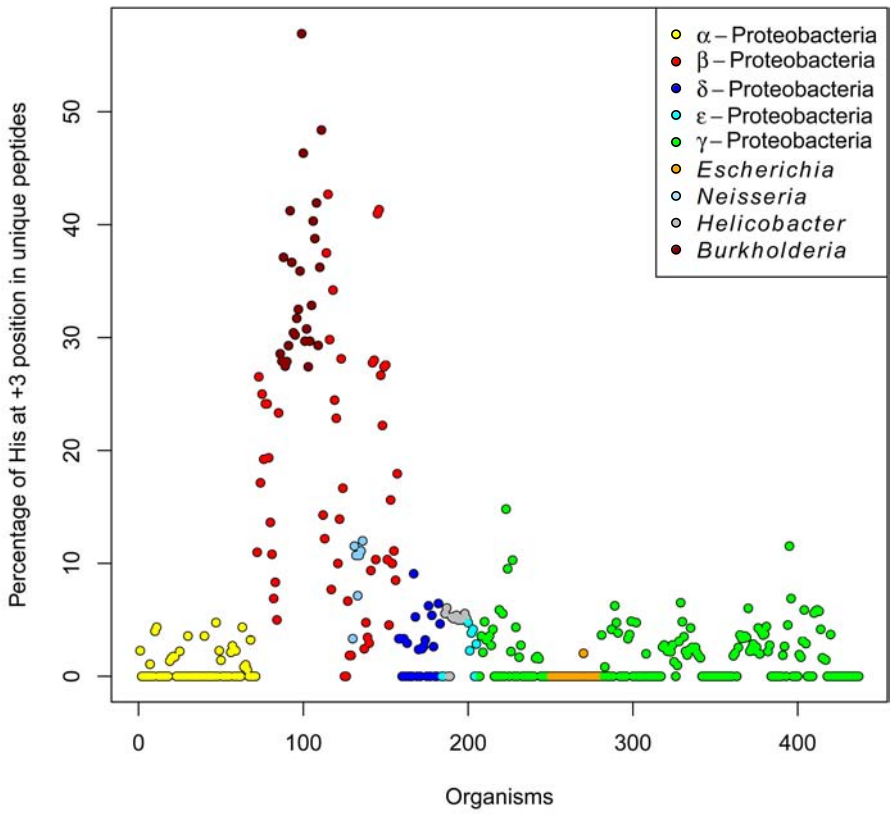

α – Proteobacteria
β – Proteobacteria
δ – Proteobacteria
ε – Proteobacteria
γ – Proteobacteria
Escherichia
Neisseria
Helicobacter
Burkholderia
Percentage of His at +3 position in unique peptides
0
10
20
30
40
50
0
100
200
300
400
Organisms

Figure 6

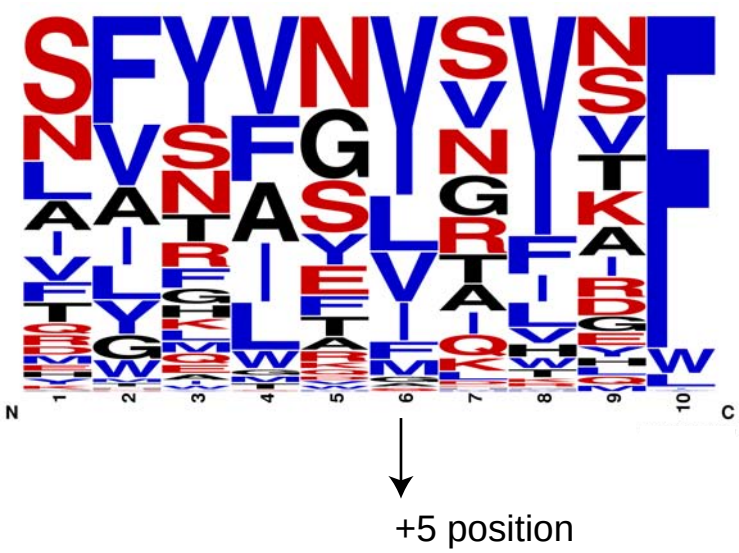
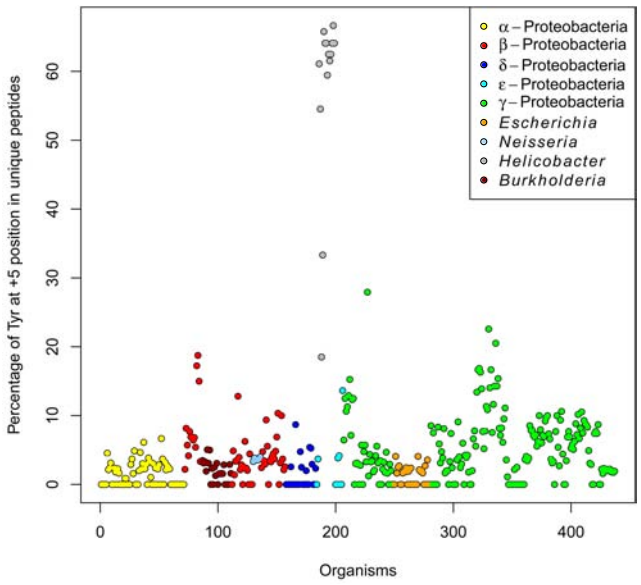


Figure 7

A


α – Proteobacteria
β – Proteobacteria
δ – Proteobacteria
ε – Proteobacteria
γ – Proteobacteria
Escherichia
Neisseria
Helicobacter
Burkholderia
Percentage of Tyr at +5 position in unique peptides
0
10
20
30
40
50
60
0
100
200
300
400
Organisms

B

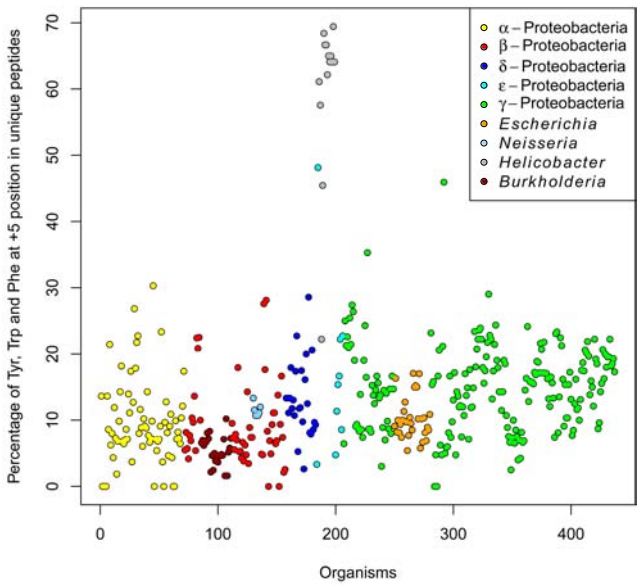

α – Proteobacteria
β – Proteobacteria
δ – Proteobacteria
ε – Proteobacteria
γ – Proteobacteria
Escherichia
Neisseria
Helicobacter
Burkholderia
Percentage of Tyr, Trp and Phe at +5 position in unique peptides
0
10
20
30
40
50
60
70
0
100
200
300
400
Organisms

Figure 8

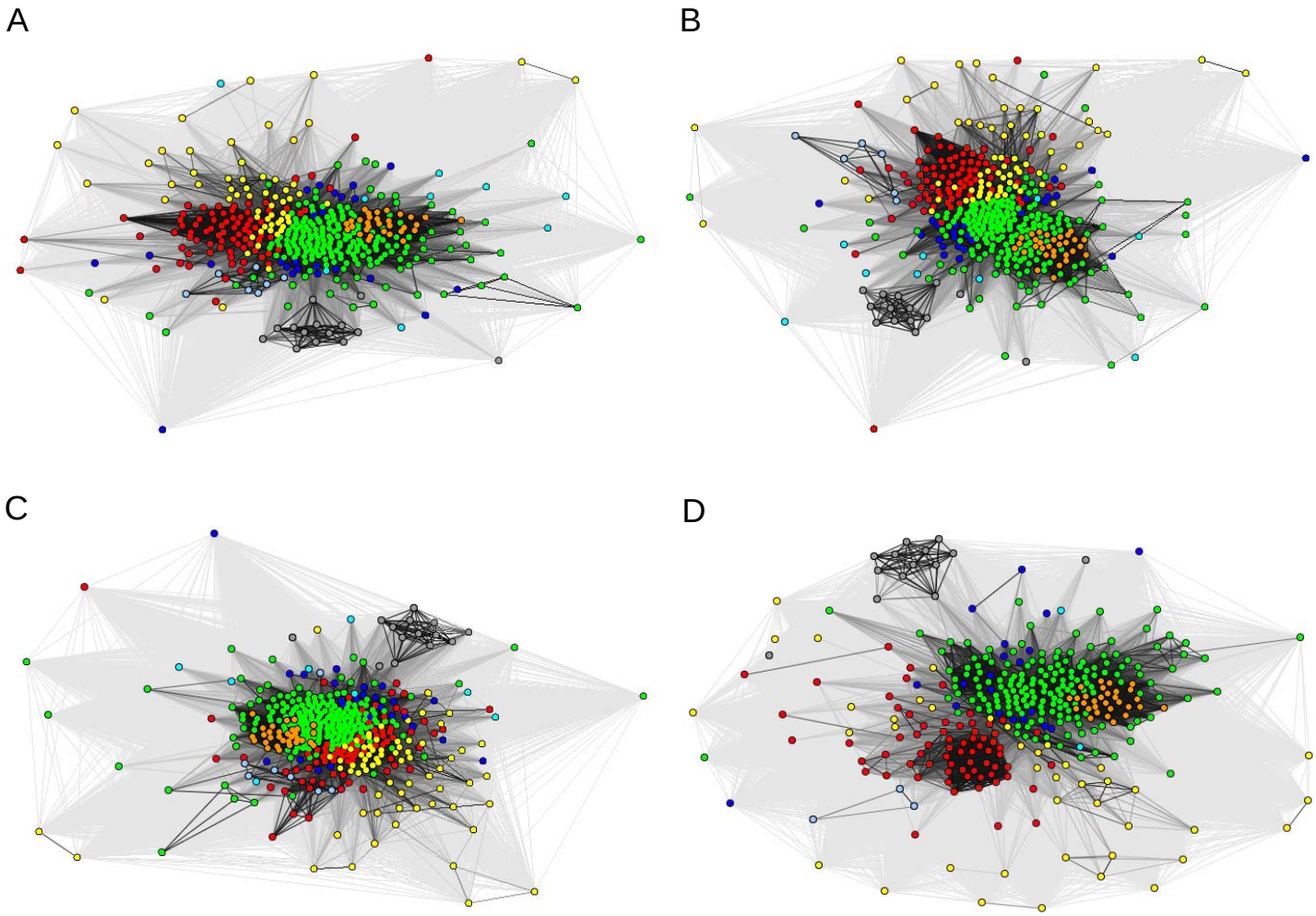


Figure 9

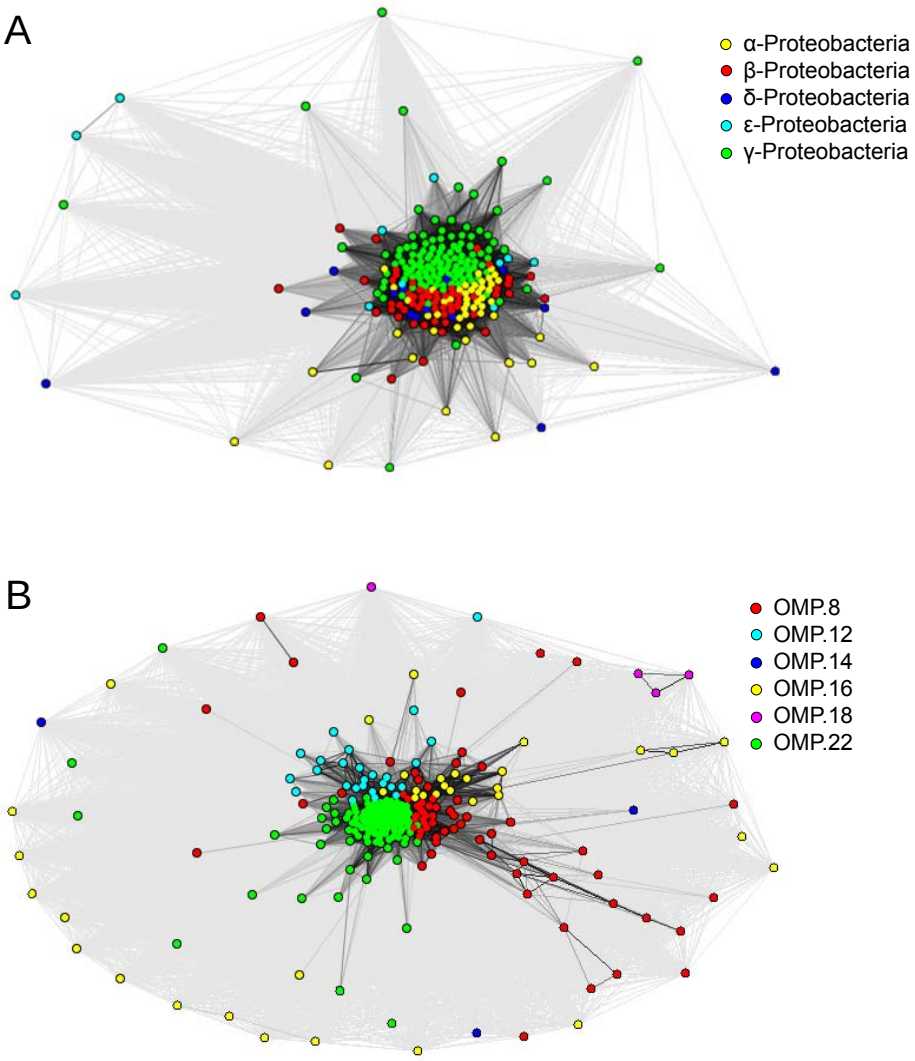

A
α-Proteobacteria
β-Proteobacteria
δ-Proteobacteria
ε-Proteobacteria
γ-Proteobacteria
B
OMP.8
OMP.12
OMP.14
OMP.16
OMP.18
OMP.22

Figure 10

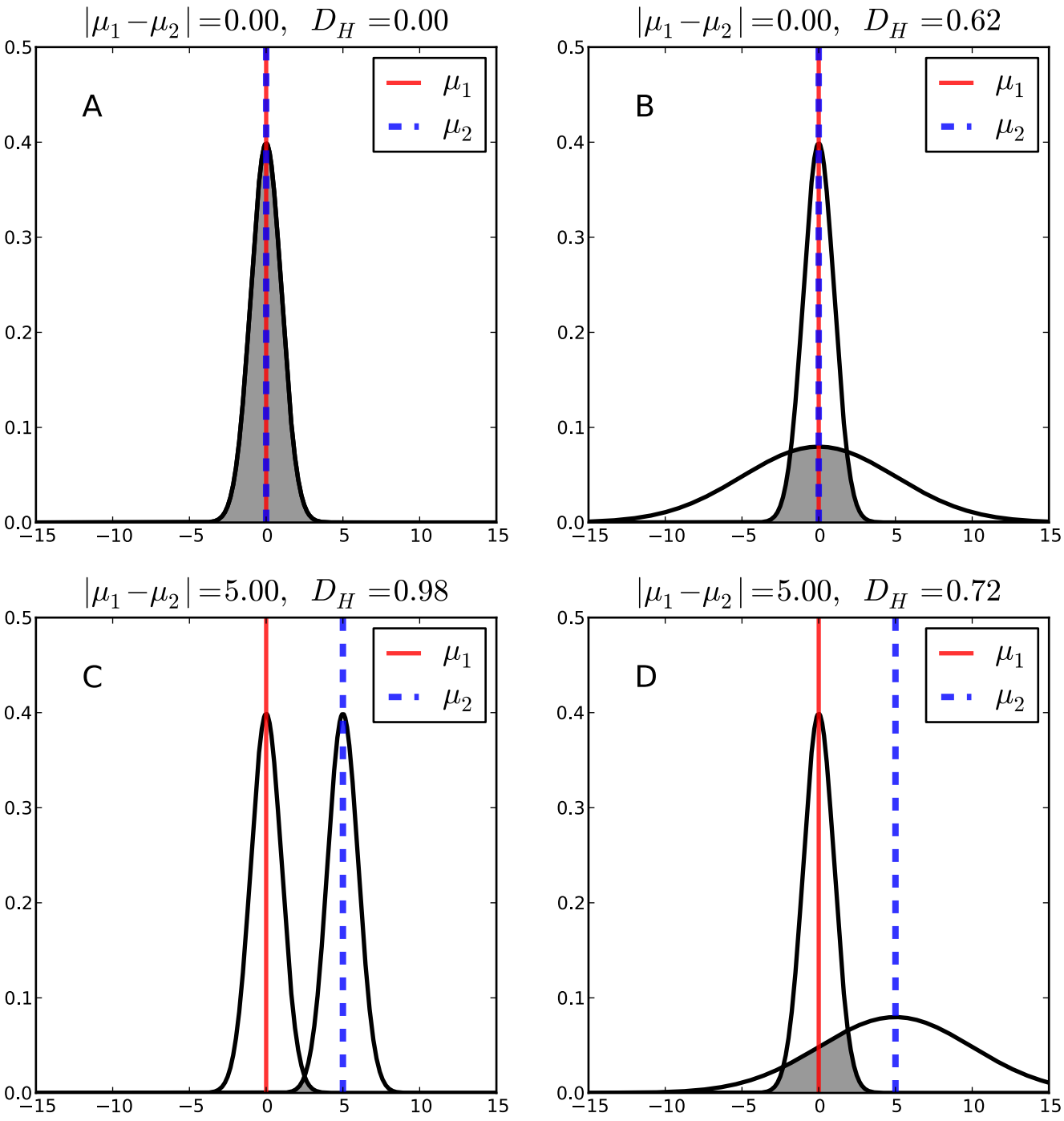
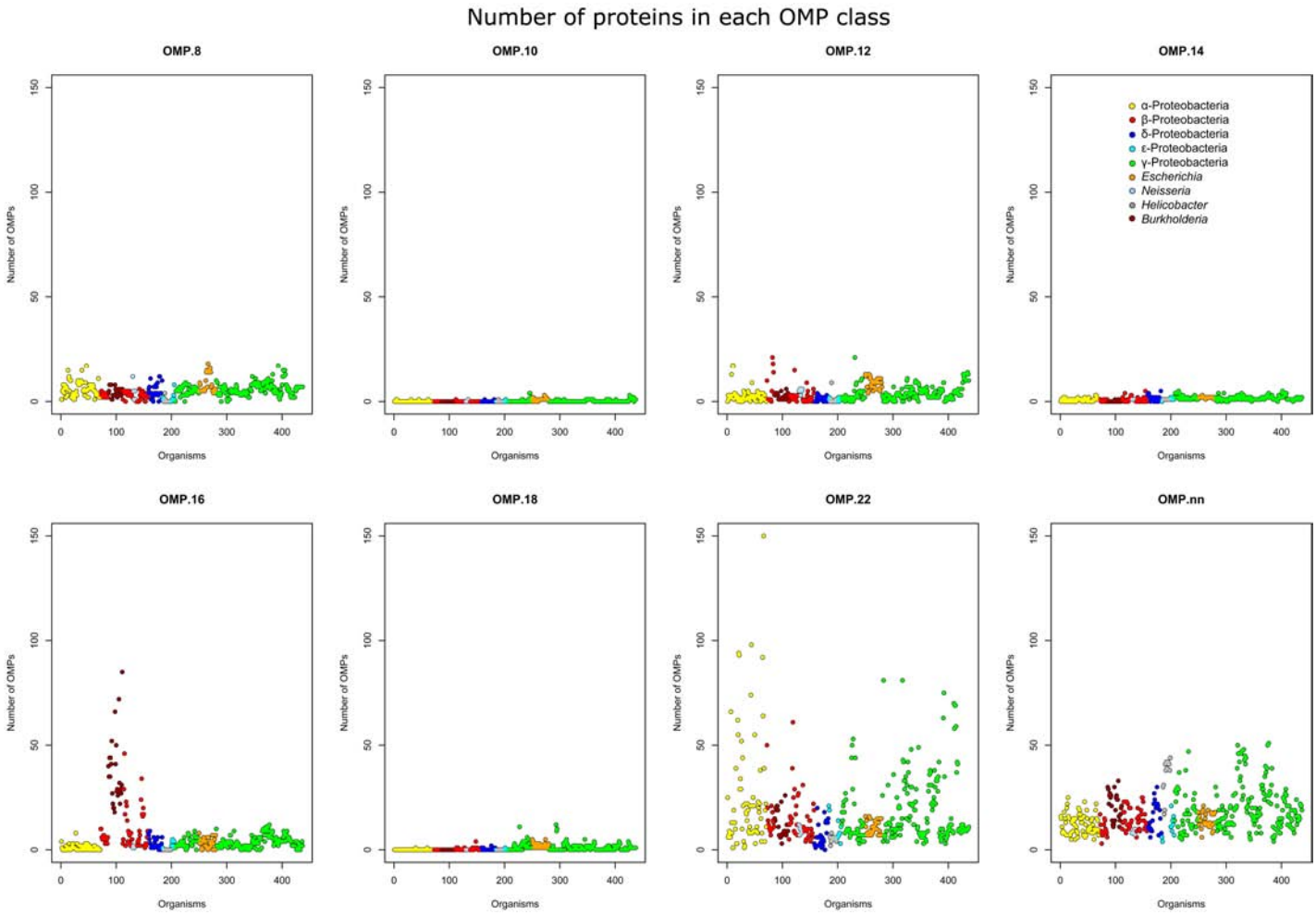

$|\mu_1 - \mu_2| = 0.00, \quad D_H = 0.00$
A
μ_1
μ_2
$|\mu_1 - \mu_2| = 0.00, \quad D_H = 0.62$
B
μ_1
μ_2
$|\mu_1 - \mu_2| = 5.00, \quad D_H = 0.98$
C
μ_1
μ_2
$|\mu_1 - \mu_2| = 5.00, \quad D_H = 0.72$
D
μ_1
μ_2
0.0
0.1
0.2
0.3
0.4
0.5
−15
−10
−5
0
5
10
15

Figure 11



Additonal file 2