UPPSALA
UNIVERSITET

UPTEC X 16 030

Examensarbete 30 hp
Oktober 2016

Bioinformatic Analysis of Genomic and Proteomic Data from Gemmata

Karl Dyrhage

UPPSALA
UNIVERSITET

Degree Project in Bioinformatics

Masters Programme in Molecular Biotechnology Engineering,
Uppsala University School of Engineering

UPTEC X 16 030 | **Date of issue 2016-10**

Author

Karl Dyrhage

Title (English)

Bioinformatic Analysis of Genomic and Proteomic data from *Gemmata*

Title (Swedish)

Abstract

Members of the bacterial phylum Planctomycetes have been claimed to have a compartmentalised cell plan, with cell walls lacking peptidoglycan despite being free-living. These theories have been challenged in recent years, and the nature of the planctomycete cell structure is currently under debate. Yet it remains clear that the planctomycete membranes have unique properties, and are thus likely localisations of evolutional innovation. In this study, proteomes and genomes of four planctomycete species from the *Gemmata*/*Tuwongella* clade were investigated with the aim to find candidate genes for functional characterisation. Analysis based on full genome sequencing and mass spectrometry revealed 21 proteins unique to the *Gemmata*/*Tuwongella* clade that were present in the proteomes of all four species. The gene coding for one of these was found to be organised in an operon, containing an additional four clade-specific genes, likely related to type II secretion. A planctomycete-specific cell surface signal peptide previously not seen in *Gemmata* was identified in all four species, with proteins found to have the motif indicating that their cell surface has a strong negative charge. Lastly, the study has revealed evidence suggesting that the planctomycetes have a traditional gram-negative cell wall, contradicting the previously proposed proteinaceous cell wall model.

Keywords
Planctomycetes, Gemmata, proteomics, subcellular localisation, functional prediction, signal peptide

Supervisors

Siv Andersson
Uppsala University

Scientific reviewer

Bengt Persson
Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **41**

Biology Education Centre Biomedical Center Husargatan 3, Uppsala
Box 592, S-751 24 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 471 4687

Populärvetenskaplig sammanfattning

Bioinformatic analysis of genomic and proteomic data from *Gemmata*

Karl Dyrhage

Planctomyceter, inklusive dess undergrupp *Gemmata*, är en bakteriegrupp som har uppmärksammats för sina ovanliga egenskaper. Bland annat har det påståtts att deras celler är uppdelade och har fler slutna utrymmen i jämförelse med andra bakterier. Denna teori, liksom andra teorier om planctomyceternas ovanliga egenskaper, är i dagsläget under debatt och kan ses som kontroversiell. Detta innebär att planctomyceterna fortfarande är en relativt okänd grupp, där mycket nytt finns att utforska.

I detta arbete har jag undersökt och jämfört de uppsättningar proteiner som finns tillgängliga och uttryckta i fyra planctomycetarter, varav tre är av släktet *Gemmata* och en *Tuwongella*, i syfte att hitta proteiner som kan vara intressanta att studera vidare i framtiden. Studien är dels baserad på bakteriernas genom, det vill säga alla de proteinkodande generna i bakterien, och dels på deras proteom, det vill säga de proteiner som har identifierats experimentellt med hjälp av masspektrometri.

Av de proteiner som här hittats i alla fyra proteom, så är det 21 som inte hittats varken i proteom eller genom hos någon annan organism utanför den studerade gruppen, och med andra ord är helt nya. Detta gör dem intressanta för vidare studier, då de kan ha tidigare osedda funktioner eller ge oss en inblick i planctomyceternas roll i naturen. Ett utav dessa proteiner har undersökts noggrannare, och föreslås vara del utav ett transportsystem. I studien läggs även fram indikationer på att planctomyceternas cellvägg har en traditionell gramnegativ struktur, något som länge betraktats som falskt.

Contents

Acronyms

BOMP	the ß-barrel outer membrane protein predictor
COG	cluster of orthogonal groups
GRAVY	grand average of hydropathy
HMM	hidden Markov model
IM	inner membrane
IMP	inner membrane protein
KEGG	Kyoto encyclopedia of genes and genomes
LC-MS/MS	liquid chromatography tandem mass spectrometry
OM	outer membrane
OMP	outer membrane protein
pI	isoelectric point
PSM	peptide spectrum match
SVM	support vector machine
TMD	transmembrane domain

1. Introduction

The planctomycetes are a largely unstudied phylum comprising environmental bacteria living in various habitats. The current state of planctomycete research is riddled with controversies, such as the claim that their cell walls lack the otherwise near-universal Gram-negative cell wall component peptidoglycan, or the claim that they have a compartmentalised cell plan. These and other claims are still being investigated, and thus the planctomycetes present a frontier with many unresolved theories, and much left to discover. This and other qualities makes them good candidate models for studying microbial evolution, and for the identification and characterisation of previously unknown genes.

While the compartmentalised cell theory remains controversial, there is little doubt that the planctomycete membranes have unusual properties. To the best of my knowledge, there have been no studies that have been published to date that utilise a large-scale proteomics approach to investigating the planctomycete membrane structure. Instead, previous large-scale studies have focused on genomic analysis. One potential problem with predicting genes from genomic data, is that many of the identified genes might not be expressed under normal conditions, or might be pseudogenes that fill no function. This problem can be bypassed by working with proteomic data. The project presented here aims to just that, and to shed new light on the planctomycetes membranes from a novel perspective, in addition to presenting general characterisation of four species of planctomycetes.

The goal of this report is to present basic, explorative research. Not much is currently known about the planctomycetes, and so the societal impact to which their characterisation may lead is difficult to predict. Nevertheless, basic science paves the way for the formulation of novel theories and the possibility of serendipity, and constitutes the foundation upon which applied science is built.

2. Background

2.1 Planctomycetes

The planctomycetes are a phylum of Gram-negative bacteria, a member of which was first described in 1924 by Hungarian scientist Nándor Gimesi [1]. The name stems from the original belief that it was a fungus. Since then many other planctomycete species have been identified, and our understanding of the phylum has improved. Many planctomycetes have long stalks that connect to those of other cells, forming small colonies. Outer membranes (OMs) of most planctomycetes also contain crateriform structures – large crater-like recessions in the membrane – that are sometimes concentrated towards one pole of the cell [2]. Other previous studies have indicated a compartmentalised cell plan present in all planctomycetes [3], otherwise almost unheard of amongst Bacteria, leading to theories that the planctomycete cell compartmentalisation might have a common origin with that of the Eukaryotes [4]. Recently this idea has been challenged by evidence from three-dimensional reconstruction of the planctomycete *Gemmata obscuriglobus* [5]. The reconstructions show that what was previously considered compartments appear to be large invaginations of the inner membrane (IM), with no closed compartments.

Another controversial claim that has been made about the planctomycetes is that all members lack peptidoglycan, an otherwise universal component of the Gram-negative cell wall, and that they have a proteinaceous cell wall instead of the normal asymmetric bilayer outer membrane [6, 7]. This would mean that they are an exception outside of the Gram-negative/positive categorisation. This idea was furthered despite evidence of genes involved in peptidoglycan synthesis in the genomes of several planctomycetes [8, 9]. The claim has been used to justify the cell compartmentalisation theory, with the argument that a peptidoglycan-free cell wall in a traditional Gram-negative bacteria would be too fragile to stay intact in the environments most planctomycetes inhabit [10]. However, newer studies have since been able to experimentally verify the presence of peptidoglycan in some species of planctomycetes [11].

Whilst the above-mentioned traits are being disputed, planctomycetes do appear to have some unusual features. The purpose and evolutionary origin of the invaginations of the IM remain a mystery. Planctomycetes also reproduce through budding rather than the standard cell fission displayed by other Bacteria [2, 12]. This, again, makes them similar to Eukaryotes, as budding is also the preferred method of division for yeast. On the genomic level, this unconventional method of reproduction is evidenced by the lack of a FtsZ homolog [2], a central protein for cell division in other bacteria. There is also evidence that some planctomycete species utilise endocytosis to take up nutrients [13], which is the first time such a mechanism has been observed in prokaryotes.

Experiments have shown that some planctomycete species are unusually resistant to exposure to UV-light [14], suggesting the presence of a sophisticated DNA repair and protection mechanisms. This makes them potential model organisms for research related to DNA decay and aging. A study investigating the presence of different bacterial phyla in various habitats found that the planctomycetes were the third most abundant marine phylum, indicating their importance as environmental bacteria **??**.

2.1.1 Gemmata

Members of the genus *Gemmata* differ from other planctomycetes in their structure, with globular cells and a uniform distribution of crateriform structures on their cell wall [15]. Their DNA is tightly packed, and clearly visible on EM images [3] (Fig. 2.1). The first *Gemmata* species to be discovered was *G. obscuriglobus*, which was isolated from a freshwater dam in Australia [15].

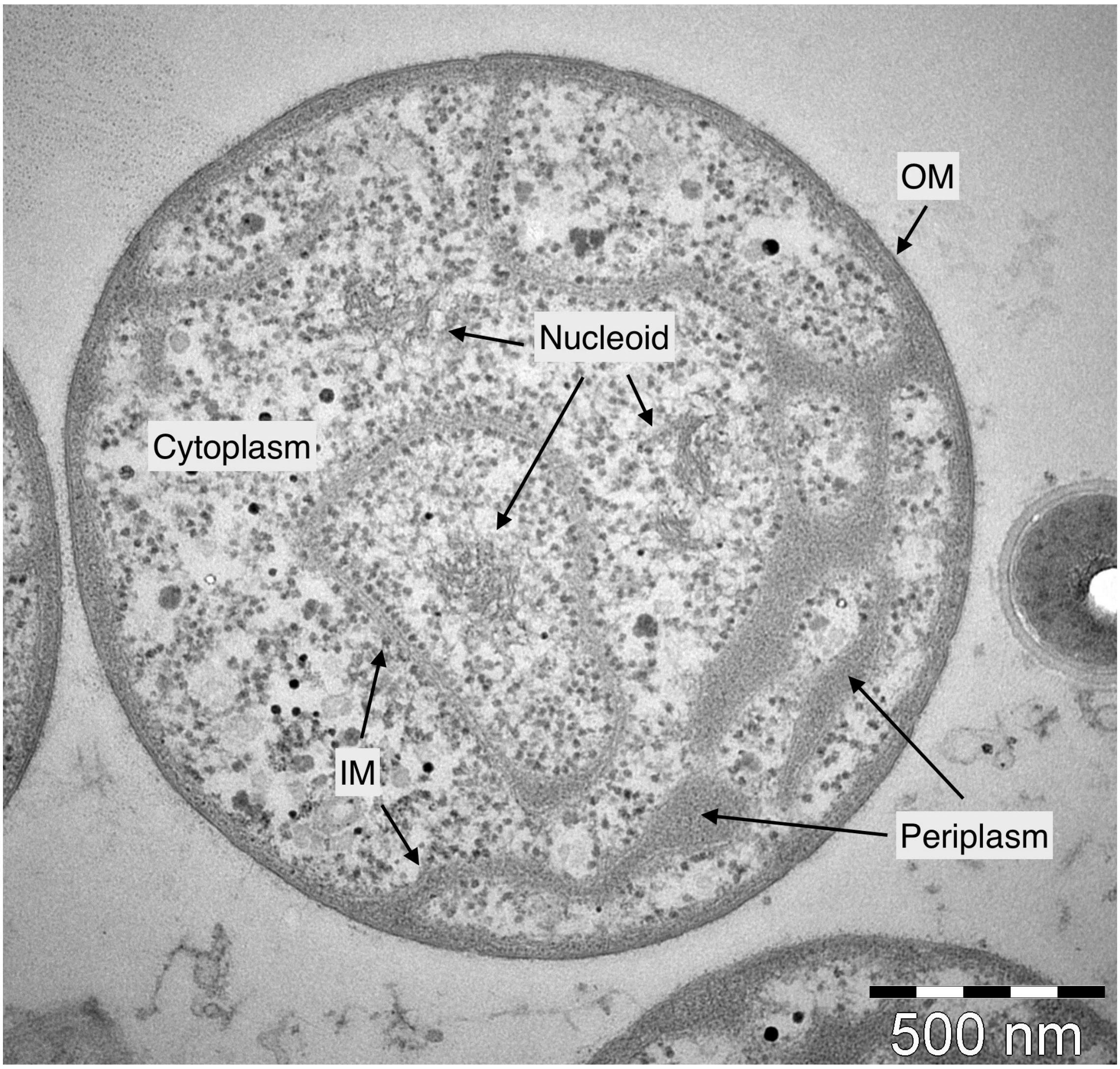


Figure 2.1. Electron microscopy image of *G. obscuriglobus*. Provided by Christian Seeger.

More *Gemmata* species have since been found in other environments, such as in waste water and in soil [16, 17]. The genome for *G. obscuriglobus* has recently been sequenced, along with two unnamed species, here referred to as GCJuql4 and GSoil9 [unpublished].

2.1.2 Tuwongella

The genome of the newly described planctomycete *Tuwongella immotidiffusa* has recently been sequenced [unpublished]. Genomic analysis has revealed that *T. immotidiffusa* is closely related to the *Gemmata*, as shown in Fig. 2.2, but phenotypical differences such as uncondensed nucleoids and non-motility suggests that it should be considered a separate genus. Due to its relatively short generation time and general ease of handling in the lab, it has potential as a model organism for studying planctomycetes.

2.2 Subcellular localisation

In Eukaryotes, the presence of organelles and specialised cell compartments is followed by a need for complex systems of protein transport and localisation. Bacteria are structurally much less complex, but they still require systems for regulating the subcellular localisation of proteins. In

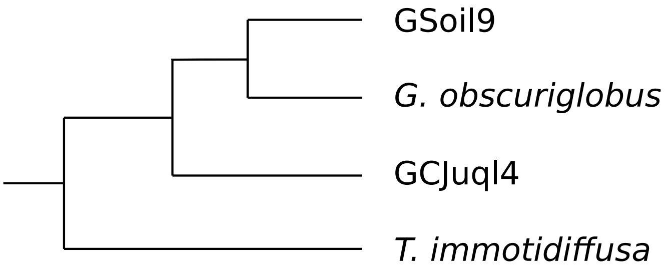


Figure 2.2. Phylogeny of the *Gemmata* and *Tuwongella*. Not to scale.

Gram-negative bacteria, the main targets for localisation, or lack thereof, are the cytoplasm, the inner membrane, the periplasmic space, and the outer membrane. While not strictly subcellular localisation, proteins can also be secreted into the extracellular space. Since the cellular structure of the planctomycetes is unclear, it should be noted that, unless specifically stated otherwise, the classical Gram-negative structure is assumed from here on.

Cytoplasm

The cytoplasm is the large innermost compartment of the bacterial cell. Proteins in the cytoplasm are mainly hydrophilic. The cytoplasm is where the DNA is located, and thus it also contains many DNA-associated proteins. Ribosomes, while often associated with the IM, are soluble and found in the cytoplasm. Protein synthesis occurs in the cytoplasm, and those that are to be localised somewhere else need to be targeted by a transportation system such as those described below.

Inner membrane

The IM, also called the cytoplasmic membrane, consists of a phospholipid bilayer, and separates the cytoplasm from the periplasm. Proteins embedded in the often have one or more hydrophobic transmembrane domains (TMDs). Another type of inner membrane protein (IMP) is the lipoprotein. Lipoproteins have in common that they have had a lipid covalently attached to them post-translationally, which allows them to be incorporated into a membrane.

Periplasm

The periplasm is similar to the cytoplasm in chemical composition. It also contains the peptidoglycan that makes up the Gram-negative cell wall. Proteins located in the periplasm must first cross the IM, using any of the multiple strategies for transporting proteins across membranes that bacteria utilise. One such system is the Tat system, which allows fully folded proteins to cross the membrane [18]. Also the Sec system can transfer proteins from the cytoplasm to the periplasm [19]. Since periplasmic proteins are hydrophilic, the signal peptides targeted by the translocation machineries are the most notable feature that sets them apart from cytoplasmic proteins.

The section of the cell referred to as the paryphoplasm in the compartmentalised planctomycete theory corresponds to the periplasm in the classical Gram-negative cell structure model [3].

Outer membrane

The OM differs from the IM in composition, as it consists of a periplasm-facing phospholipid layer, and an outermost lipopolysaccharide layer. A subgroup of outer membrane proteins (OMPs) are cell surface proteins, which are connected to or interact with the OM, while being predominantly located, or having their main function take place in, the extracellular space. Like the IM, the OM can also contain lipoproteins. The lipoproteins are first synthesised, in the cytoplasm or periplasm, and inserted into the IM, and are then transported to the OM by the Lol lipoprotein translocase system [20].

Extracellular space

Proteins that are excreted from the cell, passing both the inner and outer membranes, fall into the extracellular category. This could also include cell surface proteins which are only loosely bound to the membrane, without being embedded inside it.

2.2.1 Protein transclocation machineries

To ensure that proteins are transported to the correct location, bacteria need to utilise multiple systems for protein translocation. The Sec, Tat, and Lol systems mentioned above, are well-known examples of systems involved in the transport of proteins in Gram-negative bacteria. Proteins targeted by the Sec system have a signal peptide near the N-terminus, which is recognised by the transporter protein SecB in the cytoplasm, which binds to the polypeptide as it is being translated. It then directs it to the membrane-bound Sec translocase machinery [21, 19, 22]. From there it can be incorporated into the membrane via the insertase YidC, or sent into the periplasmic space. Similarly, the Tat system recognises a highly conserved signal peptide at the N-terminus, except only after translation is finished and the protein is fully folded, and exports the protein to the periplasm [18]. One type of proteins that is targeted by the Sec system is lipoproteins, which are inserted into the IM upon traversing the membrane. They can then be moved to the OM by the Lol system. The Lol system is comprised of a protein complex anchored in the IM which recognises and catches lipoproteins in the IM with a signal peptide, a periplasmic protein which picks up the proteins and transfers them to the OM, and an OM-bound protein which finally inserts the protein into the OM, facing the periplasmic space [23]. The lipoprotein can then be flipped to the extracellular side of the OM by the LptDE protein complex [24].

3. Data

3.1 Genomic data

Genome assemblies for *T. immotidiffusa* and three species of *Gemmata* (*G. obscuriglobus*, GCJuql4, GSoil9) were provided by Mayank Mahajan, along with genes predicted from open reading frames, with BLAST-based annotations.

3.2 Clade-specific proteins

A set of 149 proteins specific to the *Gemmata*/*Tuwongella* clade were provided by Mayank Mahajan. Proteins were assigned to this set if they, based on OrthoMCL clustering, are present in all four of the species used in this study, and no orthologs were found in any other organism. OrthoMCL clusters are groups of putative orthologs identified using the OrthoMCL software [25].

3.3 Mass spectrometry data

Liquid chromatography tandem mass spectrometry (LC-MS/MS) data was provided for *T. immotidiffusa* and the three *Gemmata* strains by Christian Seeger. Each organism was analysed with three biological replicates. From each replicate, proteins for which $\geq$2 peptides were found with $\geq$95% confidence were combined to form a final list of experimentally verified proteomes. Provided with the list of proteins is the sum of the peptide spectrum matches (PSMs) for a protein from all three replicates. The PSM is the number of peptides found in the LC-MS/MS experiment that map to that protein, which can be used as a rough estimate of abundance.

For *T. immotidiffusa* only, LC-MS/MS data from fractionated proteome experiments was available. Data from one experiment was available initially, and two more, using different fractionation protocols, were added throughout the duration of the project. Only data from the final iteration is reported here. The experiments resulted in three fractions, S1, S2, and S3, with the purpose of enriching IMPs in the second fraction, S2.

To briefly summarise the protocol used for the fractionation, the cells were cultivated in medium for 68 hours, and lysed using sonication. The resulting solution was centrifuged and separated into the supernatant (S1), which was saved, and the pellet, which was resuspended in a solution containing Triton X-100, a hydrophobic surfactant, to extract IMPs. The resulting solution was then centrifuged, separated into supernatant (S2) and pellet. The new pellet was resuspended in an SDS-contatining solution, centrifuged, and the supernatant (S3) was saved (Table. 3.1). Similar fractionation protocols have been proven successful in other Gram-negative bacteria, such as *E. coli* [26, 27].

Table 3.1. Summary of fractions extracted from *T. immotidiffusa* for fractionated LC-MS/MS, with the solution each fraction was suspended in, and the type of proteins expected to be enriched in that fraction.

Fraction	Solution	Content
S1	Tris	Soluble proteins
S2	Tris + Triton X-100	Cytoplasmic membrane
S3	Tris + SDS	Outer membrane

Table 3.2. Number of proteins identified in LC-MS/MS experiments for four planctomycetes.

Species	Replicate 1	Replicate 2	Replicate 3	Combined
T. immotidiffusa	1201	1119	1167	1565
G. obscuriglobus	1265	1352	1409	1554
GCJuql4	1293	1246	1266	1476
GSoil9	770	792	744	887

Table 3.3. Number of proteins identified in fractionated LC-MS/MS experiments for *T. immotidiffusa*

Experiment #	Fraction 1	Fraction 2	Fraction 3
1	172	1047	1014
2	642	1099	1686
3	1092	973	1433

4. Methods

4.1 Subcellular localisation prediction

Various general and specialised softwares were used for subcellular localisation prediction

PSORTb

PSORTb [28] (version 3.0.2) is a general subcellular localisation predictor. It has separate modes for archaea, Gram-negative, and Gram-positive bacteria. In Gram-negative bacteria, it assigns each investigated protein to one of the following categories:

- Cytoplasmic
- CytoplasmicMembrane
- Periplasmic
- OuterMembrane
- Extracellular
- Unknown

The prediction is based on several internal predictors, such as support vector machines (SVMs) trained on labelled data from each of the subcellular localisations, signal peptide prediction, and SubCellular Localisation-BLAST, which runs the investigated protein sequence against a database of proteins with known localisations and assigns a score based on sequence similarity. Each internal predictor produces a score signifying the likelihood of a particular localisation. Finally, a prediction is made by combining the individual results.

CELLO

CELLO [29] is a general subcellular localisation predictor. It can work with eukaryotic, Gram-negative, and Gram-positive data. The predictions are made using a two-layered SVM system, where the first layer consists of SVMs trained on data from each respective localisation, and the second layer trained using output from the first layer to make the most likely final prediction. Unlike PSORTb, CELLO has no Unknown category. Instead a prediction is forced on every protein, even if there is no strong signal.

Phobius

Phobius [30, 31] is a predictor for TMDs, using a hidden Markov model (HMM). Proteins with one or more TMDs tend to be localised in the inner membrane. It also attempts to predict signal peptides, which have similar hydrophobic qualities as TMDs and could yield false positives if not taken into account.

BOMP

The ß-barrel outer membrane protein predictor (BOMP) [32] predicts the presence of β-barrel membrane spanning structures. Proteins with β-barrels can fulfill various functions, but have in common that they are localised in the OM [33]. Thus BOMP can be used to verify predictions made by other softwares.

LipoP

LipoP [34] is a predictor for lipoproteins, that works by identifying a signal sequence located near the N-terminal. Lipoproteins get inserted into the IM. Depending on their signal sequence, they then either remain there, or are transferred to the outer membrane by the Lol translocation machinery. Yamaguchi et al. [35] reported that the sorting of lipoproteins depends on a single amino acid located 2 residues from the cleavage site. For that reason LipoP returns the amino acid at this position. However, the lipoprotein sorting signal has since been shown to involve more of the surrounding residues [36], and thus no attempt at using this information to predict the final localisation has been made in this project.

SignalP
SignalP [37] is a predictor for proteins targeted by the Sec secretion system. It looks for a well-conserved signal peptide, that is located near the N-terminal. Identified proteins are assumed to be non-cytoplasmic.

Other
Other predictors that were used in the project, but played less central roles, were TMHMM [38], a TMD predictor based on hidden Markov models, and TatP [39], a neural network-based predictor for the Tat signal peptide.

4.2 Protein statistics

Physical properties for proteins were predicted using Pepstats [40] (molecular weight, isoelectric point, charge), and GRAVY calculator [41] (grand average of hydropathy (GRAVY) index). GRAVY index is an estimation of hydropathicity, calculated as the mean hydropathicity of the amino acids in a protein, where proteins with an index >0 are assumed to be hydrophobic, and <0 hydrophilic [42]. Proteins whose sequences contain **X** (an unspecified amino acid) cannot be processed by Pepstats, and are thus excluded from analyses using molecular weight, isoelectric point, or charge.

4.3 Functional prediction

Functional prediction was performed using the cluster of orthogonal groups (COG) and Kyoto encyclopedia of genes and genomes (KEGG) databases, as well as InterProScan [43]. InterProScan is a tool for functional analysis, which queries multiple databases such as Pfam [44], SUPERFAMILY [45], and PROSITE [46], to predict protein domains. The COG database contains protein domain motifs, which are assigned to different generalised categories (see Appendix B). A single domain may be assigned to multiple categories, and a protein may have multiple domains. The KEGG database also aims to predict high-level functionality, by classifying proteins into more specific categories such as particular metabolic pathways, based on orthology.

4.4 Operon identification

In order to investigate the function of certain proteins, an in-house script was used for checking whether a given protein is found in an operon or not. The script was written in Julia [47], a relatively new programming language for technical and scientific computing, marketed as having high performance compared the more well-established alternatives (R, MATLAB, Python, etc.).

The script takes a list of IDs for the proteins to investigate for each organism, where each list contains the proteins assigned to the same OrthoMCL clusters as the other lists, in the same order. For each protein family the script compares the OrthoMCL clusters of the surrounding genes for each organism. If all organisms have more than *N* proteins, where *N* is defined by the user, from the same set of OrthoMCL clusters, that locus is considered a potential operon. Apart from *N*, the user can also set the size of the window of surrounding proteins to compare, by defining the number of proteins upstream and downstream of the investigated protein to include, as well as the minimum number of organisms that must share a cluster for it to count as a valid hit.

4.5 Motif identification

A previous study in *Rhodopirellula baltica*, another planctomycete, identified a novel signal peptide found on the N-terminus of cell surface and extracellular proteins, but failed to identify proteins with the same motif in *G. obscuriglobus* [48]. At the time the genome of *G. obscuriglobus* had not been fully sequenced, so an attempt to identify this signal peptide in the now fully sequenced *Gemmata* genomes as a continuation of their study. This was done using a script written in the Julia programming language. As input the script takes:

`--infiles`	list of FASTA files containing the sequences of interest
`--motif`	string containing the motif to search for
`--cutoff`	maximum allowed distance from the N-terminus
`--allowedmissing`	number of allowed mismatches
`--variable`	list of positions in the motif that may take alternate forms
`--nproc`	number of additional processes to spawn, up to (# available processors)-1

as well as options related to output. Running the script with default settings on the four planctomycete genomes takes 3 minutes and 11 seconds. With `--nproc 7` the time is reduced to 45 seconds. The program writes the results to two files: one file containing the identified sequences aligned around the motif in FASTA format, and one containing the distance from the N-terminus and the number of mismatches for each identified sequence. Using the above described script and the motif described in *R. baltica* as a starting point, I investigated the predicted proteomes of all four planctomycetes, and cross-referenced the results with prediction data for subcellular localisation and physical properties.

5. Results

5.1 Hydrophobicity and isoelectric points

Schwartz et al. [49] showed that the proteomes of bacteria and eukaryotes display bimodal and trimodal isoelectric point (pI) distributions, respectively, and proposed that this is a result of the complexity of the eukaryotic cell, where proteins in different subcellular localisations are exposed to different conditions. Based on this, if the planctomycetes have compartmentalised cells like eukaryotes, we would expect to observe similar trimodal distributions. To verify this, pI was predicted for proteins from all four planctomycetes and *E. coli* using the ExPASy server [50]. All displayed bimodal distributions, where the three *Gemmata* species had pI distributions similar to each other, having their highest peaks >7, whereas *T. immotidiffusa* and *E. coli* had their highest peaks <7 (Fig. 5.1A). Extending the hypothesis above to the hydropathicity of membrane proteins, which may be under different conditions if there are multiple membranes within the cell, GRAVY index was calculated using GRAVY Calculator [41]. The GRAVY index distributions were more similar within the planctomycetes, with *E. coli* being noticeably distinct from all of them being the only one with a clearly bimodal distribution (Fig. 5.1B).

5.2 Subcellular localisation prediction

The genomically inferred proteomes for all four planctomycetes, as well as *E. coli*, were used as input for PSORTb and CELLO, both set to Gram-negative. The two gave the same predictions for 76-78% of all proteins, depending on the organism being investigated, when excluding those annotated as Unknown by PSORTb. While the genome size differs between the four planctomycete species, the ratios of proteins predicted in each category are very similar between species, as shown in Fig. 5.2. The LC-MS/MS data contained 1565, 1554, 1477, and 887 proteins for *T. immotidiffusa*, *G. obscuriglobus*, GCJuql4, and GSoil9, respectively. This represents 10-30%

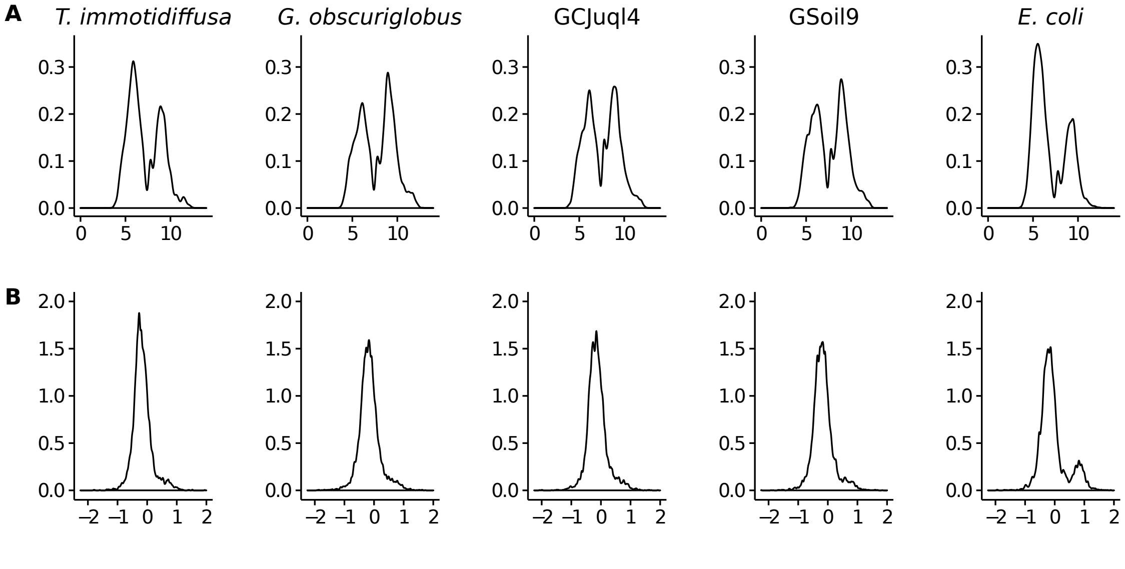


Figure 5.1. Distribution of proteins in *T. immotidiffusa*, three species of *Gemmata*, and *E. coli*, for (A) isoelectric point, and (B) GRAVY index.

Table 5.1. Number of proteins predicted to have a given localisation by PSORTb.

Prediction	*T. immotidiffusa*	*G. obscuriglobus*	GCJuql4	GSoil9	*E. coli*
Cytoplasmic	1894	2413	2203	2870	1946
Extracellular	39	63	44	85	47
Inner Membrane	1038	1251	1036	1399	1077
Outer Membrane	34	47	37	59	91
Periplasmic	94	214	226	224	161
Unknown	2134	3602	2972	3940	948
Total	5233	7590	6518	8577	4270

Table 5.2. Number of ribosomal proteins in proteomic / genomic data.

Subunit	*T. immotidiffusa*	*G. obscuriglobus*	GCJuql4	GSoil9
30s	22 / 24	24 / 29	20 / 20	20 / 26
50s	27 / 31	24 / 25	28 / 29	25 / 31

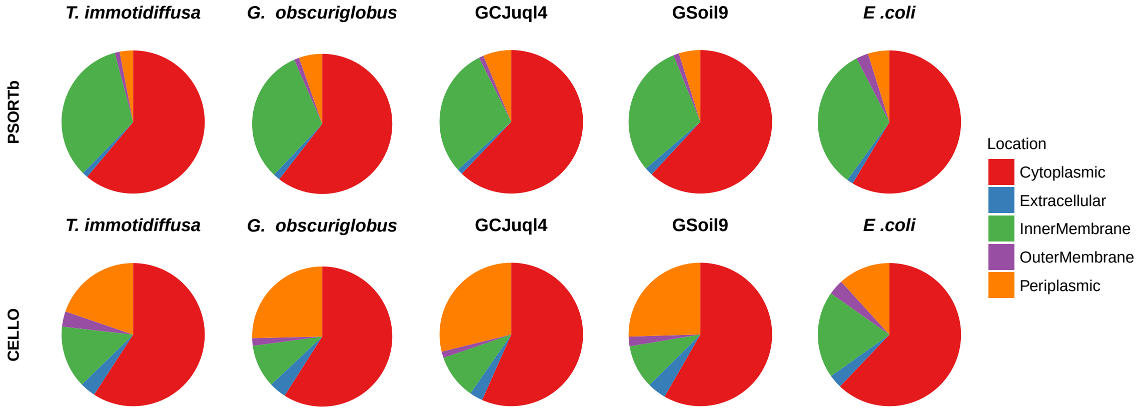


Figure 5.2. Pie charts summarising subcellular localisation predictions for genomic data, from PSORTb and CELLO. Proteins that were considered Unknown by PSORTb were excluded from that analysis.

of the total proteomes inferred from the genomic data (Table 5.1). Unlike the total numbers of proteins identified in the proteomes of the four species, the ratios of proteins in each localisation is nearly constant between them (Fig. 5.3). It is notable that cytoplasmic and periplasmic proteins were overrepresented in the proteomics data, mainly at the expense of IMPs. An analysis based on annotations showed that the majority of all proteins from the small and large ribosomal subunits were covered in the proteomes of all four organisms (Table 5.2), a possible indication that the identified sets of proteins are representative of the actual expressed proteome.

To verify that the fractionated LC-MS/MS experiment had resulted in an enrichment of IMPs in the second fraction, I compared the relative abundances, i.e. the PSM of a protein in each fraction normalised by the combined PSM of that protein, in the three fractions of proteins from each respective PSORTb localisation. Proteins annotated as IMPs were mainly found in S2 as expected (Fig. 5.4). This was particularly noticeable when looking at proteins exclusive to that fraction, where 79% (50% when including those categorised as Unknown) were predicted IMPs. For the proteins annotated as IMPs that were present in but not exclusive to S2, there was a positive correlation between the relative abundance and both the hydrophobicity and the number of TMDs (Fig. 5.5). The same did not hold true for those annotated as cytoplasmic.

I manually compiled a list of 104 proteins with known subcellular localisations that were present in the fractionated experiment, using a combination of literature review, genomic annotation, and sequence similarity with known proteins in *E. coli*. Proteins in this list include members of the Sec [19], Tat [18], and Lol [20] systems for IMPs, phospholipases and lysophospholipases for OMPs [51], and ribosomal proteins and tRNA ligases for cytoplasmic proteins (see Appendix A for the full list). These proteins were then used to further evaluate the results of the fractionated LC-

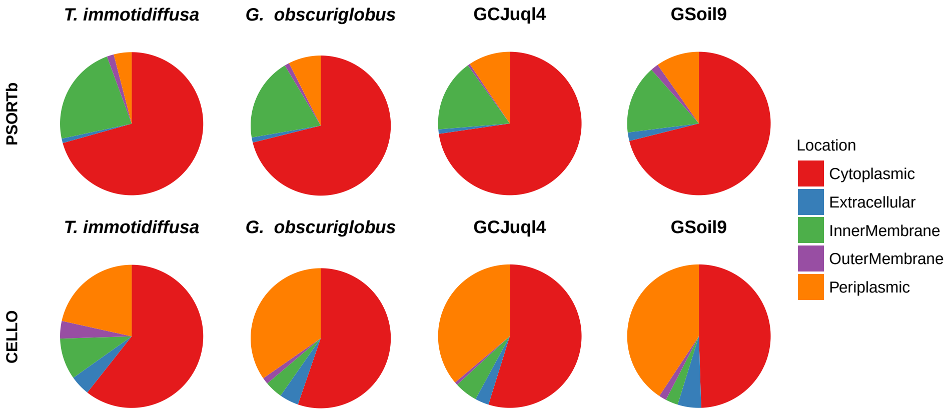


Figure 5.3. Pie charts summarising subcellular localisation predictions for proteomic data verified through LC-MS/MS, from PSORTb and CELLO. Proteins that were considered Unknown by PSORTb were excluded from that analysis.

MS/MS experiments for *T. immotidiffusa*. Most IMPs and OMPs were found mainly or exclusively in S2 and S3, respectively, while the cytoplasmic proteins were found in both S1 and S3. The only periplasmic protein that was included was spread between all three fractions (Fig. 5.6).

5.3 Core proteome

The core proteome was inferred by finding orthologous proteins, based on OrthoMCL clustering, that are present in the genomes and proteomes of all four species. 2391 proteins were found to be shared by all species, out of which 471 were also identified with LC-MS/MS in all for species (Fig. 5.7A).

5.4 Clade-specific proteins

Among the 149 *Gemmata*/*Tuwongella* specific proteins, 21 proteins were found to be experimentally identified in all four species (Fig. 5.7C). 20 of these were also found in the fractionated MS experiment in *T. immotidiffusa*. Three of these were found mainly in S1, 13 in S2, as well as one exclusively in S3, while three were spread out between multiple fractions (Fig. 5.8).

Of the 13 found predominantly in S2 there were 11 that were predicted to have TMDs according to Phobius, and one that was predicted to have a lipoprotein signal peptide by LipoP. The protein not found in any fraction also had a lipoprotein signal peptide. This means that out of the set of 21 proteins, 12 are both predicted to be membrane-bound and are found in the IMP-enriched LC-MS/MS fraction, and two proteins were either predicted to be membrane-bound, or found in S2, giving a total of 67% IMPs. Compared to the highest estimate of the fraction of IMPs in the genome, about 33% (Fig. 5.2), IMPs are greatly overrepresented in this set.

The protein found in S3 had a Sec signal peptide. The equivalent proteins in the three *Gemmata* species were found to have multiple YTV repeat domains by InterProScan, a domain previously described in *R. baltica* OMPs [52]. Upon aligning the sequences from all four species, manual inspection revealed the same repeats in *T. immotidiffusa*. These three observations taken together make it likely that it is an OMP.

5.5 Functional analysis

All sequences were queried against the COG database to find conserved domains. 58%, 52%, 56%, and 53% of the proteins from *T. immotidiffusa*, *G. obscuriglobus*, GCJuql4, and GSoil9, respectively, had at least one COG domain, compared to 85% for *E. coli*.

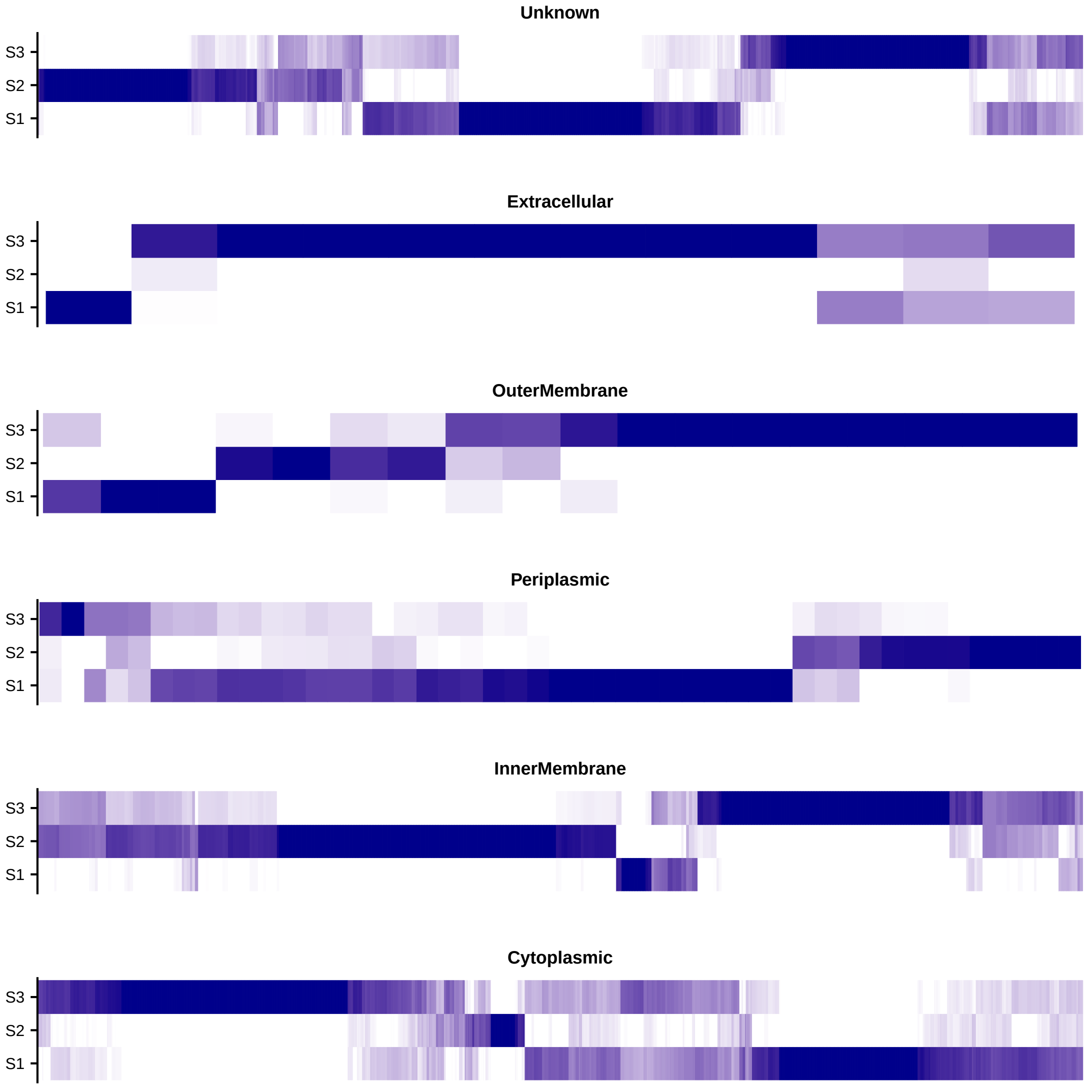


Figure 5.4. Relative abundance of proteins from fractionated LC-MS/MS experiment in *T. immotidiffusa*, grouped by localisation predicted by PSORTb. Proteins were ordered horizontally using hierarchical clustering.

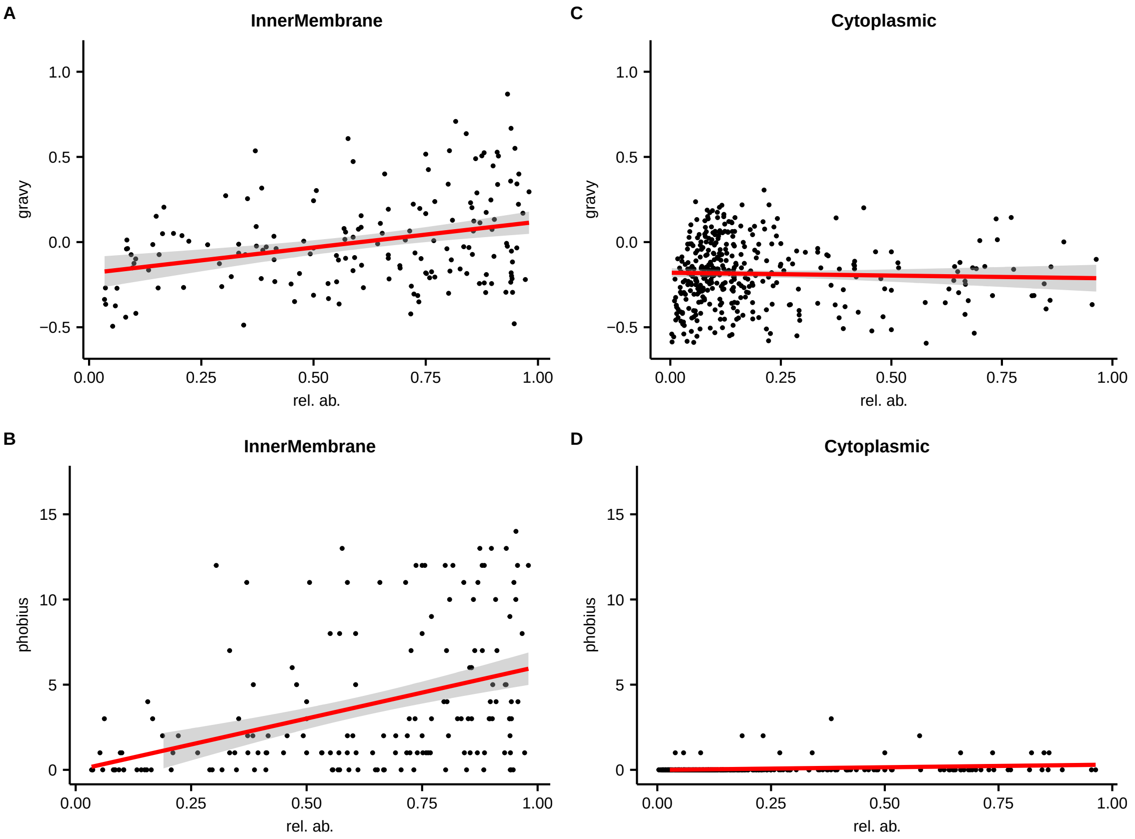


Figure 5.5. Relative abundance in IMP-enriched LC-MS/MS fraction for proteins annotated as IM/C by PSORTb, against (A-B) GRAVY index, and (C-D) number of TMDs identified by Phobius. Only proteins found in multiple fractions are included. Black dots represent individual proteins, and red lines represent linear models of the displayed data.

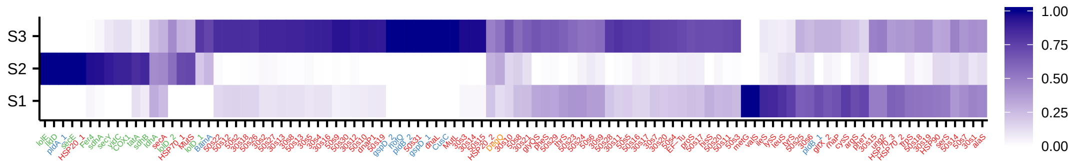


Figure 5.6. Relative abundance of proteins from fractionated LC-MS/MS experiment in *T. immotidiffusa* for proteins with known localisation, sorted by hierarchical clustering. The scores were normalised by the sum of each row. Protein names are coloured according to expected localisation, where red represents C, green IM, blue OM, and orange P.

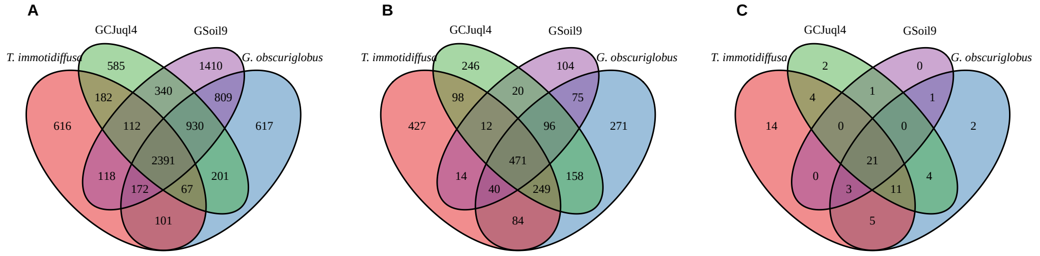


Figure 5.7. Venn diagram showing overlap of orthologous proteins present in (A) the full genomic sets of proteins, (B) proteins found with LC-MS/MS, and (C) proteins unique to the *Gemmata*/*Tuwongella* clade and part of their core proteome. Orthology is based on OrthoMCL clusters.

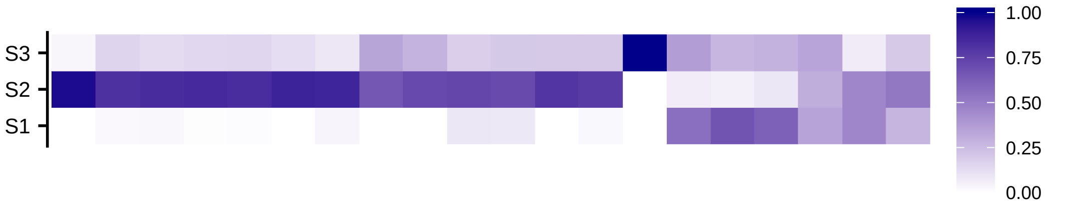


Figure 5.8. Relative abundance of proteins from fractionated LC-MS/MS experiment in *T. immotidiffusa* for the proteins unique to the *Gemmata*/*Tuwongella* clade and experimentally verified in all four species.

I compared the percentage of COGs found in each category, for the genomic and proteomic data of the planctomycetes. All showed very similar patterns, with COG domains related to energy production, amino acid metabolism, carbohydrate metabolism, translation, and post-translational modification being overrepresented in the proteomic data in all four organisms. Similarly, replication and repair were underrepresented in all proteomes. Similar patterns were observed for the proteins comprising the core proteome (Fig. 5.9).

For the second fraction of the fractionated *T. immotidiffusa* LC-MS/MS data, energy production and conversion and defense mechanisms were the two most overrepresented categories compared to the two other fractions. Replication and repair, signal transduction, and coenzyme transport and metabolism, on the other hand, were underrepresented (Fig. 5.10).

KEGG annotations for all four species from GhostKOALA and the KEGG Automatic Annotation Server were provided by Mayank Mahajan. I compared the number of proteins from different KEGG modules and pathways present in the genome and proteome between the different species, to find an explanation for the reduced proteome size in GSoil9. The citrate cycle, pyruvate metabolism, glycolysis, and carbon metabolism pathways are notable examples where all four had similar numbers of proteins in the genome, and most were found in the proteome. For both purine and pyrimidine metabolism, however, the GSoil9 proteome contained between one third to half of the amount of proteins in the same categories in the other species (data not shown). Also notable is that the three *Gemmata* genomes all contained 23 to 24 proteins related to flagellar assembly, whereas *T. immotidiffusa* only has one.

5.6 Operon search

The operon search script described in the methods section was used to investigate whether any of the 21 *Gemmata*/*Tuwongella*-specific identified in all four proteomes were organised in operons on the genomes. When looking at up to 10 genes up- and downstream of the query genes, I found two potential operons where all four species shared 10 and 13 genes, respectively.

In addition to the gene from the set of 21 proteins used to identify the operon, the 10-gene operon encodes for four more clade-specific proteins, each found experimentally in three out of four species. Analysis with InterProScan showed that the first four genes encode for proteins related to type II secretion. The following five, which includes the additional clade-specific genes, contain type IV pilin N-term methylation sites. One of the clade-specific genes was present in two copies in GCJuql4. The tenth gene was the one used to find the operon, and has yielded no hits with any database. All proteins from the operon were found experimentally in at least three species (Fig. 5.11). The four proteins related to type II secretion were all found mainly in S3 in the fractionated LC-MS/MS experiment, while the remaining six were found mainly in S2 (data not shown). The 13-gene operon did not contain any additional clade-specific genes apart from the query gene. Analysis with InterProScan did not uncover any clear patterns that could hint at the function of the query gene (Appendix C).

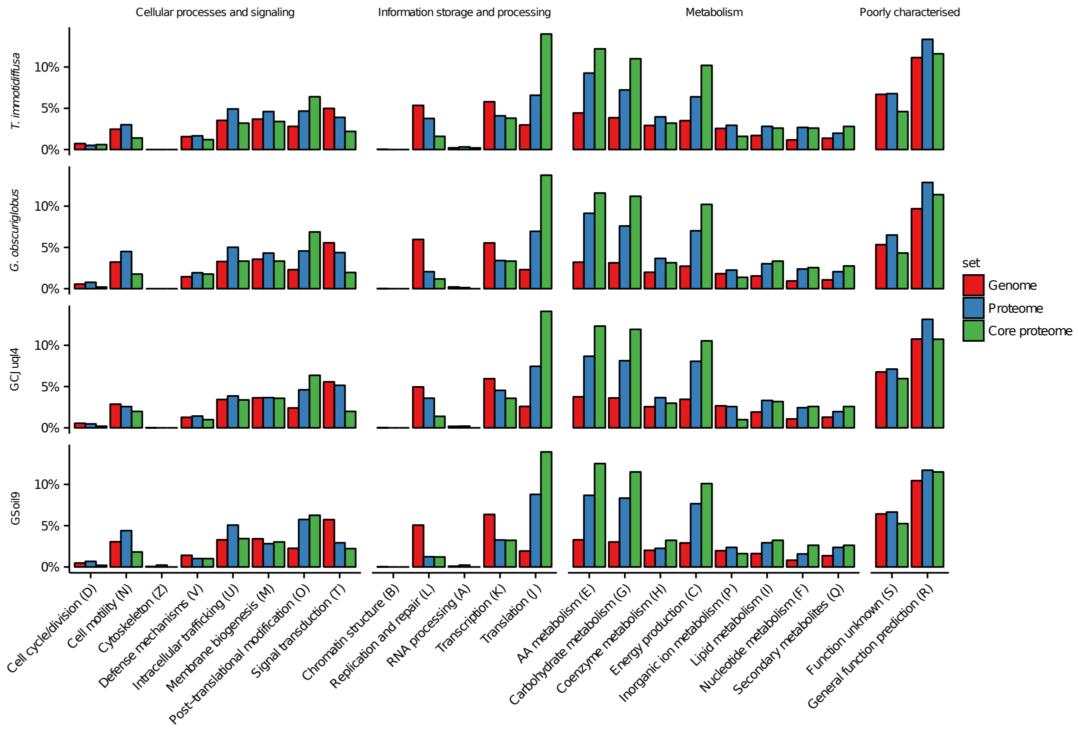


Figure 5.9. Bar plot showing the percentage of proteins with a COG domain in a given category in (red) the whole genome, (blue) the LC-MS/MS verified proteome, and (green) the core proteome, shared by all species based on OrthoMCL clustering.

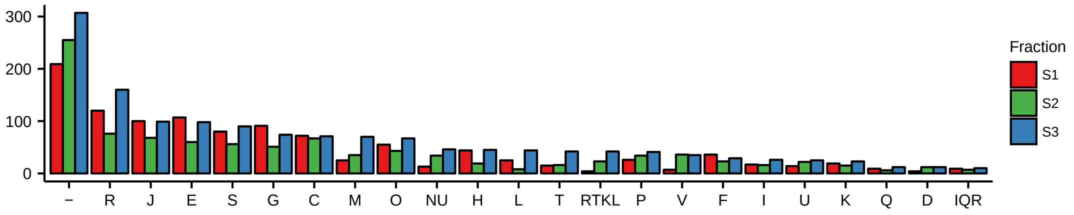


Figure 5.10. Bar plot showing the absolute number of proteins with a COG domain in a given category, in each LC-MS/MS fraction for *T. immotidiffusa*. Combinations of categories represent protein domains assigned to multiple functional categories.

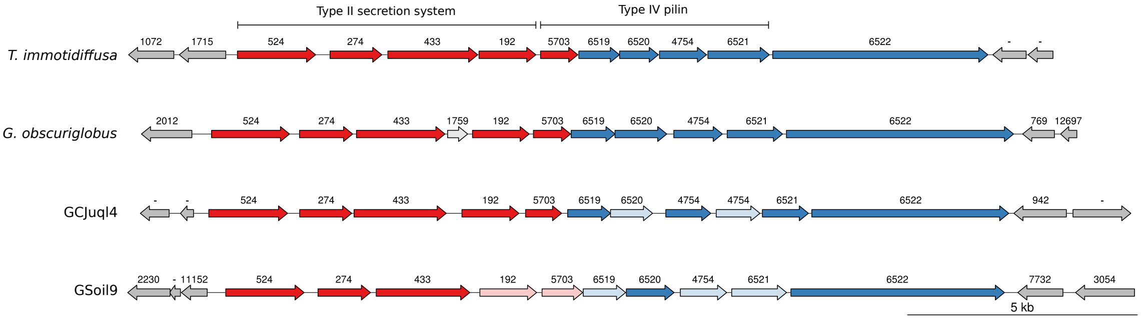


Figure 5.11. Genomic map of an operon containing genes unique to the *Gemmata*/*Tuwongella* clade. Coloured arrows represent genes found in all four species, with blue representing those unique to the clade. Arrows with strong colour were identified experimentally in that species. Numbers above arrows show the OrthoMCL cluster ID for that protein. The functional annotations shown above *T. immotidiffusa* are based on InterProScan. Created using genoPlotR.

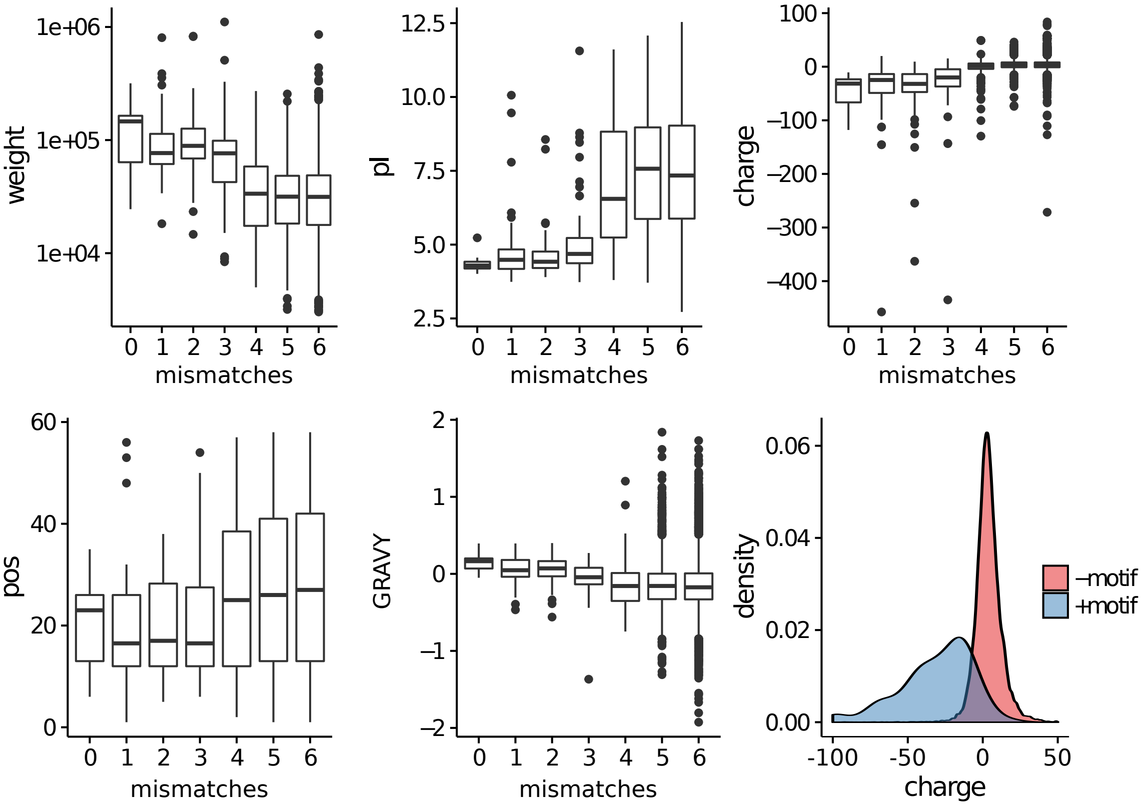


Figure 5.12. Box plots showing (A) molecular weight, (B) isoelectric point, (C) charge, (D) GRAVY index, and (E) motif start position for proteins found to have a cell surface signal peptide motif with up to six mismatches, and (F) a density plot of the charge for proteins with and without the signal peptide. Proteins are considered to have the motif it is present with three or fewer mismatches up to 70 residues away from the N-terminus.

5.7 Motif search

Using the script described in the methods section, the *Gemmata*/*Tuwongella* equivalent of the signal peptide previously described in *R. baltica* by Studholme et al. (2004) was identified. The motif that gave the best results, visualised as a sequence logo in Appendix D, can be represented as **Lx[VL]ExLEDRx[VT]PA**. Proteins identified with this motif near the N-terminus tend to be larger than average, have low isoelectric points, be hydrophobic, and have negative charge. These tendencies remain with up to three mismatches (Fig. 5.12). When allowing three mismatches, *T. immotidiffusa*, *G. obscuriglobus*, GCJuql4, and GSoil9 are found to have 52, 49, 36, and 48 proteins with the motif, respectively. 94% of all proteins identified in this manner are predicted to be either extracellular or outer membrane proteins by CELLO. For PSORTb it is 33%, although when only considering proteins with a prediction other than Unknown it jumps up to 82%. 37% were found to have β-barrels by BOMP, a very high portion compared to 1% for the full genomes. Of the 28 identified proteins that were detected in the fractionation experiment, 26 were found mainly in S3, and the remaining two were found partly in S3.

As the motif is said to be unique to the planctomycete clade, any occurrences in other organisms should be accidental. Therefore the *E. coli* proteome was used as a reference to crudely investigate the possibility of finding the motif by chance. Using the same settings used to identify the motif in *Gemmata*/*Tuwongella*, with up to 3 mismatches, yielded zero matches. Compared to the 36-52 proteins found in the planctomycetes, it appears unlikely that they were found by coincidence.

6. Discussion

6.1 Subcellular localisation prediction

An important aspect to consider regarding localisation prediction is that in order to perform it, we have to make an assumption about the cell structure of the organism under investigation. The planctomycetes have previously been described as having a cell plan that is very different from that of conventional Gram-negative bacteria, with a peptidoglycan-free proteinaceous cell wall and closed compartments within the cell. If that were the case, they would need specialised systems for protein translocation, which would likely diverge from those found in most Gram-negative bacteria. Some forms of localisation prediction, such as TMD prediction, would not be affected by this, but predictions that rely on identifying signal peptides would. The signal peptide for outer membrane/extracellular localisation is an example where we know that the planctomycetes differ from other phyla. On the other hand variation in signal peptides is not uncommon; it has for example been shown that there can be a correlation between GC bias and the signal peptide sequences utilised by an organism [53]. Most recent evidence suggests that the planctomycetes do have a regular Gram-negative cell wall, making it reasonable to assume that they use the same systems for protein translocation. Whether the inner structure of the planctomycete cell is compartmentalised or just invaginated, an interesting project for the future would be to search for innovation or specialisation in the translocation machineries. This could for example be done by employing the same methods used to identify the novel signal peptide in *R. baltica* to the newly sequenced planctomycetes, or using experimental methods for investigating localisation such as GFP tagging.

The two generalised subcellular localisation predictors that were used in this study, PSORTb and CELLO, have one major difference; while PSORTb has an Unknown category, CELLO gives a prediction for every analysed protein even if there is no significant signal. As can be seen in Fig. 5.2 CELLO predicts a much larger fraction of the proteins to be periplasmic: as many as 1033 compared to 94 periplasmic proteins predicted by PSORTb if *T. immotidiffusa*. According to Weiner and Li (2007), the size of the *E. coli* periplasmic proteome has been estimated to lie between 141 and 367 proteins, where the higher predicton was criticised for including many known membrane-associated proteins. The same study also mentioned that at least 124 proteins have been experimentally verified to be periplasmic. When comparing the predictions for *T. immotidiffusa*, which have genome sizes comparable to those of *E. coli*, with these numbers, it seems likely that the set of 1033 proteins predicted by CELLO contains a large number of false positives.

A different case where the difference between the two predictors can be observed is in the set of proteins identified to have the cell surface/extracellular signal peptide. 94% of all proteins identified to have the motif are predicted by CELLO to be localised in the OM or to be extracellular. With PSORTb, only 34% are given the same predictions, while the rest are mostly Unknown. If we assume that the signal peptide is indeed working the same way in *Tuwongella* and *Gemmata*, as in *Rhodopirellula*, then this is a case where PSORTb has a large number of false negatives.

The above examples show that PSORTb has high precision, while CELLO has high recall. In other words, the former attempts to minimise the number of false positives while the latter maximises the number of true positives. This should be taken into consideration when analysing the results. When verifying the fractionated LC-MS/MS experiment precision was of higher relevance, thus PSORTb was used. When examining proteins with the cell surface signal peptide however, PSORTb had very low recall. In that case the high recall of CELLO was more relevant. In conclusion, both predictors should have a potential place in the bioinformaticians arsenal, and the choice of which tool to employ depends on the question that is under investigation.

6.2 Fractionated LC-MS/MS experiment validation

Several of the analyses that I performed were aimed at validating the results of the fractionated LC-MS/MS experiment. The list of proteins with known localisation (Appendix A, Fig. 5.6) was one such analysis. It shows that IMPs and OMPs are enriched in S2 and S3, as expected, while cytoplasmic proteins are spread out between S1 and S3.

The proteins that were found in S2 while also being predicted to be cytoplasmic could be explained as soluble proteins that are associated with the IM. Another possible explanation is that their localisations are falsely predicted. However, as seen in Fig. 5.5, cytoplasmic proteins found in S2 do not display the same correlation between relative abundance in the fraction, and hydrophobicity and number of TMDs, as does predicted IMPs. If they were true IMPs falsely predicted as cytoplasmic, they should follow the same pattern. Additionally, as soluble proteins vastly outnumber membrane-bound proteins in the samples, there is a risk for carry-over from S1 to S2 and S3. This, again, suggests that the explanation lies in biochemistry rather than in informatics.

6.3 Proteins of special interest

Of the 149 protein families that were present in the genomes of all four *Gemmata*/*Tuwongella* species, and no other, 21 were also found in the proteomes of all four. These should be good targets for experimental characterisation if the goal is to look for innovations among these organisms. As shown in section 5.6, one of the 21 was found together with four other clade-specific proteins in a potential operon. While not found experimentally in all species, all four of the additional clade-specific proteins in the operon were found in three out of four species, which is a strong indication that this operon is, in addition to being conserved between all four species, highly expressed and of functional importance. Functional analysis revealed that four out of ten proteins in the operon contained domains related to type II secretion systems, which transfer substrates across the OM. These four were all found experimentally in S3, suggesting that they could be OM-associated. Another five proteins contained type IV pilin signal peptides, a motif sometimes found in proteins related to type II secretion [54]. The tenth protein, which was the query protein used to find the operon, was found in high abundance in S2. Together these observations suggest that the operon is related to type II secretion, possibly with the query protein anchored in the IM, the four type II secretion proteins anchored in the OM, and the five type IV pilin signal peptide bearing proteins functioning as a bridge between them.

6.4 Cell surface signal peptide

The signal peptide motif was identified based on that described in *R. baltica* [48]. While only the central residues of the motif appear to be conserved between *Rhodopirellula* and the *Gemmata*/*Tuwongella* clade, the sets of proteins identified with the motif share some physiological properties. Studholme et al. (2004) [48] describe them as being larger on average compared to the whole set of proteins in *R. baltica*, and that functional prediction found many proteins with domains related to the cell surface and extracellular functions. The size observation holds true in the species that I have investigated, and most of the protein domains mentioned in the *R. baltica* article are are found as well. From this it is clear that it is indeed the same signal peptide described here. In addition to the trend in protein size, I also found a strong bias towards negative charge and low isoelectric points. For continued studies on this subject, it would be interesting to see whether the same patterns can be found in *R. baltica*, and indeed other planctomycetes.

Apart from the non-core sequence, one aspect that differs from *R. baltica* is the distance from the N-terminus at which the motif is found. In *Rhodopirellula*, the first residue of the motif (counted as the first leucine) has a mean position of 49.5 and a median of 47, compared to 22 and 18 for *Gemmata*/*Tuwongella*. In fact, the earliest occurrence of the motif in *Rhodopirellula* is at position

18 [48]. These differences are not wholly unexpected, as signal peptides have been shown to vary between species [53].

6.5 Phylogeny and phenotype

The genomes of *G. obscuriglobus* and GSoil9 contain the largest numbers of genes of the four species (Table 5.1). They also share the greatest number of proteins of any two species (Fig. 5.7), in concurrence with the phylogeny portrayed in Fig. 2.2. Similarly, the greatest number of proteins shared by any three species is between the three *Gemmata*. Due to the small size of the GSoil9 proteome compared to the other taxa, it is difficult to do a similar analysis based on proteomic data. With that in consideration, however, we can look at the number of proteins shared between the GSoil9 proteome and each of the other three species. As expected, the number is the highest for GSoil9/*G. obscuriglobus*, with 75 compared to 20 and 15 for GSoil9/GCJuql4 and GSoil9/*T. immotidiffusa*. The proteome of *T. immotidiffusa* also contained the largest number of proteins only found that species, with 427 compared to 104-271 for the *Gemmata*, which again supports the phylogeny.

One phenotypical feature that sets *T. immotidiffusa* apart from the *Gemmata*, as evidenced in its name, is its immotility [unpublished]. This is supported by the KEGG analysis, which showed that the *T. immotidiffusa* genome lacks genes related to flagellar assembly, unlike the *Gemmata* which all had 23-24 such genes. The other part of its name, *diffusa*, refers to the fact that its nucleoid is more loosely packed than those of *Gemmata*, and not as clearly visible with electron microscopy. I did not look for evidence of this in the proteomes, but it is likely that if one were to look for proteins related to DNA organisation a similar pattern would emerge.

6.6 Experimental and bioinformatic analysis

While previous bioinformatic studies of planctomycetes have focused on genomic analysis, which in the era of next generation sequencing could be considered nearly purely bioinformatic tasks, this project has contained both computational and experimental aspects. While my own contribution has been solely of the former nature, it has depended on the success of the latter. Likewise, the enhancements that were done to the fractionation protocol over the course of the project were based on the informatical analyses that attempted to verify their success. In other words, the combination creates an almost synergistic effect. Compare the 149 clade-specific that had been identified using only genomic data, with the 21 found in the core proteomes using experimental data. When searching for candidate proteins for characterisation the former number is very daunting, while the latter is much more manageable. Experimental verifaction also ensures that the proteins are truly expressed in the cells, eliminating any potential falsely predicted genes.

6.7 Summary

This project has validated the proteomic data produced for the four planctomycetes *T. immotidiffusa*, *G. obscuriglobus*, GCJuql4, and GSoil9, by showing that they were consistent between species, and contained representative core proteomes. It has shown that bioinformatic methods developed for Gram-negative bacteria yield similar results in these organisms as in classical Gram-negative bacteria. It has also validated the fractionated proteomics data, based on methods developed specifically for bacteria with an outer membrane, from *T. immotidiffusa*. Furthermore, it has presented experimental evidence of the presence of outer membrane marker proteins, such as the outer membrane protein assembly factor BamA, in planctomycete proteomes. These observations suggest that the plancomycete cell wall is more similar to the traditional Gram-negative cell wall than what has previously been suggested.

The project has resulted in the identification of 21 core proteins unique to the *Gemmata*/*Tuwongella* clade, and an operon coding for a putative type II secretion system, containing multiple clade-specific genes. These are prime candidates for future functional characterisation. Additionally, a planctomycete-specific cell surface signal peptide which previous attempts failed to identify in *G. obscuriglobus* was found and verified both experimentally and bioinformatically. Many of the proteins identified with the domain have unknown function, and could be interesting targets for characterisation as well.

Since the plancomycetes are highly abundant environmental bacteria, it is likely that they play important roles in the ecosystem. Characterisation of the candidate proteins identified in this study may give us a better understanding of those roles, and could help in the development of tools for battling evironmental pollution, for example.

Acknowledgements

I would like to thank Siv Andersson for allowing me to work on this project, and for the numerous opportunities at learning through experience that she and her lab have provided me in the past, without which I would not be the bioinformatician that I am today. I would like to thank Christian Seeger, not only for being an exemplary student and picking up ggplot very quickly, but also for the advice, guidance, and discussions, and for producing much of the data that I have worked on. Similarly, I would like to give thanks to the other two members of the planctomycete group, Mayank Mahajan, and Benjamin Yee, who provided me with the data and knowledge that this project was based on.

Appendices

A. Proteins with known localisation

Accession ID	Name	Localisation	Comment
GMBLW141190	DegQ	P	Serine endoprotease
GMBLW150280	BamA	OM	Outer membrane protein assembly protein
GMBLW128650	pldA_1	OM	Phospholipase [51]
GMBLW141420	pldA_2	OM	Phospholipase [51]
GMBLW104370	pldB_1	OM	Lysophospholipase [51]
GMBLW137090	pldB_2	OM	Lysophospholipase [51]
GMBLW101560	TolC	OM	Outer membrane efflux protein
GMBLW109640	CusC	OM	Cation efflux protein
GMBLW134940	gspD_1	OM	Putative type II secretion system
GMBLW147910	gspD_2	OM	Putative type II secretion system
GMBLW137880	hofQ	OM	Putative DNA transport protein
GMBLW116790	sdhA	IM	Succinate dehydrogenase [51]
GMBLW116800	sdhB	IM	Succinate dehydrogenase [51]
GMBLW111700	Fer4	IM	Ferredoxin
GMBLW133480	lldD	IM	L-lactate dehydrogenase
GMBLW118900	ldhA	IM	D-lactate dehydrogenase [51]
GMBLW111640	COX1	IM	Cytochrome-c oxidase
GMBLW110470	tatA	IM	Tat translocase
GMBLW100520	tatC	IM	Tat translocase
GMBLW112540	secY	IM	Sec translocase
GMBLW114660	secE	IM	Sec translocase
GMBLW124590	secG_1	IM	Sec translocase
GMBLW150120	secG_2	IM	Sec translocase
GMBLW103150	yidC	IM	Inner membrane insertion protein
GMBLW135670	lolD_1	IM	Lipoprotein-releasing system ATP-binding protein
GMBLW111860	lolD_2	IM	Lipoprotein-releasing system ATP-binding protein
GMBLW135680	lolE	IM	Lipoprotein-releasing system transmembrane protein
GMBLW145070	secA	P	Sec translocase cytoplasmic component
GMBLW125610	metG	C	Methionine tRNA ligase
GMBLW125420	hisS	C	Histidine tRNA ligase
GMBLW123790	ileS	C	Isoleucine tRNA ligase
GMBLW116480	glyQS	C	Glycine tRNA ligase
GMBLW106090	proS	C	Proline tRNA ligase
GMBLW104500	gltX_1	C	Glutamate tRNA ligase
GMBLW102760	gltX_2	C	Glutamate tRNA ligase
GMBLW102690	leuS	C	Leucine tRNA ligase
GMBLW150840	valS	C	Valine tRNA ligase
GMBLW144940	trpS	C	Tryptophan tRNA ligase
GMBLW142110	tyrS	C	Tyrosine tRNA ligase
GMBLW141540	thrS	C	Threonine tRNA ligase
GMBLW136790	pheS	C	Phenylalanine tRNA ligase subunit
GMBLW133160	pheT	C	Phenylalanine tRNA ligase subunit
GMBLW136590	alaS	C	Alanine tRNA ligase
GMBLW135690	lysS	C	Lysine tRNA ligase
GMBLW119230	argS	C	Arginine tRNA ligase
GMBLW106530	serS	C	Serine tRNA ligase
GMBLW142890	cysS	C	Cysteine tRNA ligase
GMBLW141570	glutR	C	Glutamyl tRNA reductase
GMBLW116990	rplC	C	Ribosomal l3
GMBLW150010	rplM	C	Ribosomal l13
GMBLW142090	rplS	C	Ribosomal l19
GMBLW112640	rpmC	C	Ribosomal l29
GMBLW132360	rpsA	C	Ribosomal s1
GMBLW103050	rpsB	C	Ribosomal s2
GMBLW112660	rpsC	C	Ribosomal s3
GMBLW112490	rpsD	C	Ribosomal s4
GMBLW112560	rpsE	C	Ribosomal s5
GMBLW142840	rpsF	C	Ribosomal s6
GMBLW114570	rpsG	C	Ribosomal s7
GMBLW112590	rpsH	C	Ribosomal s8
GMBLW150000	rpsI	C	Ribosomal s9
GMBLW112720	rpsJ	C	Ribosomal s10
GMBLW112500	rpsK	C	Ribosomal s11
GMBLW114580	rpsL	C	Ribosomal s12
GMBLW112510	rpsM	C	Ribosomal s13
GMBLW112600	rpsN	C	Ribosomal s14
GMBLW120050	rpsO	C	Ribosomal s15
GMBLW142070	rpsP	C	Ribosomal s16
GMBLW112630	rpsQ	C	Ribosomal s17
GMBLW136480	rpsR	C	Ribosomal s18
GMBLW112680	rpsS	C	Ribosomal s19

Continued on next page

Accession ID	Name	Localisation	Comment
GMBLW103070	rpsT	C	Ribosomal s20
GMBLW115810	rpsU	C	Ribosomal s21
GMBLW114670	EF-Tu	C	Elongation Factor Tu
GMBLW100940	RecN	C	DNA repair protein
GMBLW136930	LigD	C	ATP-dependent DNA ligase
GMBLW137180	MutT	C	DNA mismatch repair protein
GMBLW122100	MutL	C	DNA mismatch repair protein
GMBLW122390	dnaL	C	DNA ligase
GMBLW136570	ung2	C	uracil-DNA glycosylase
GMBLW141950	MutS	C	DNA mismatch repair protein
GMBLW146840	AraC	C	DNA-binding domain containing protein
GMBLW108670	rnaP_1	C	RNA polymerase Sigma70_r4_2
GMBLW108660	rnaP_2	C	RNA polymerase Sigma70_r2
GMBLW108510	rnaP_3	C	RNA polymerase Sigma-E factor
GMBLW107410	dnaP1	C	DNA polymerase I
GMBLW108300	dnaP3	C	DNA polymerase III subunit delta
GMBLW118560	rnaP	C	RNA polymerase alpha subunit domain protein
GMBLW116830	HSP20_1	C	HSP 20
GMBLW116820	HSP20_2	C	HSP 20
GMBLW116550	HSP20_3	C	HSP 20
GMBLW113690	DnaJ	C	HSP, curved DNA binding protein cbpA
GMBLW132110	HSP70_1	C	HSP 70
GMBLW110030	HSP70_2	C	HSP 70
GMBLW103470	HSP70_3	C	HSP 70
GMBLW103030	HSP90	C	HSP 90
GMBLW141560	HSP	C	HSP

B. COG Guide

Key	Full name
D	Cell cycle control, cell division, chromosome partitioning
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
O	Post-translational modification, protein turnover, and chaperones
T	Signal transduction mechanisms
U	Intracellular trafficking, secretion, and vesicular transport
V	Defense mechanisms
W	Extracellular structures
Y	Nuclear structure
Z	Cytoskeleton
A	RNA processing and modification
B	Chromatin structure and dynamics
J	Translation, ribosomal structure and biogenesis
K	Transcription
L	Replication, recombination and repair
C	Energy production and conversion
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
G	Carbohydrate transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport, and catabolism
R	General function prediction only
S	Function unknown

C. 13-gene Operon

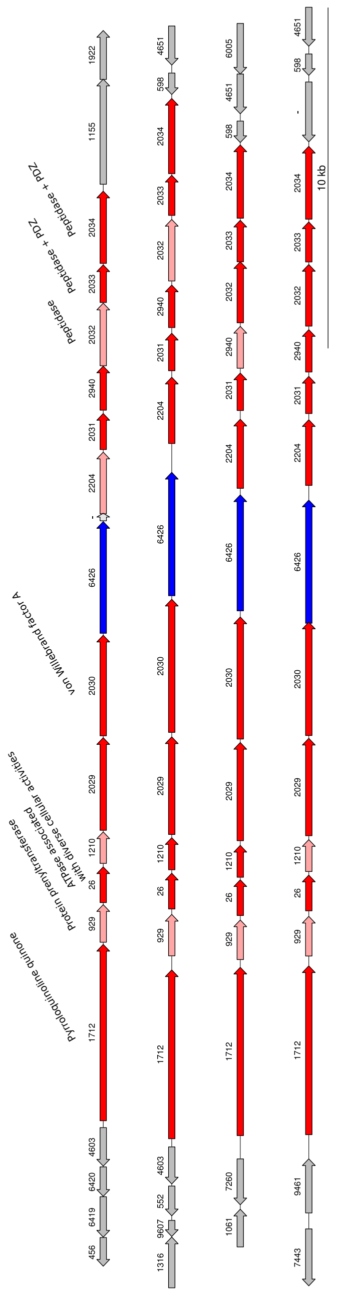
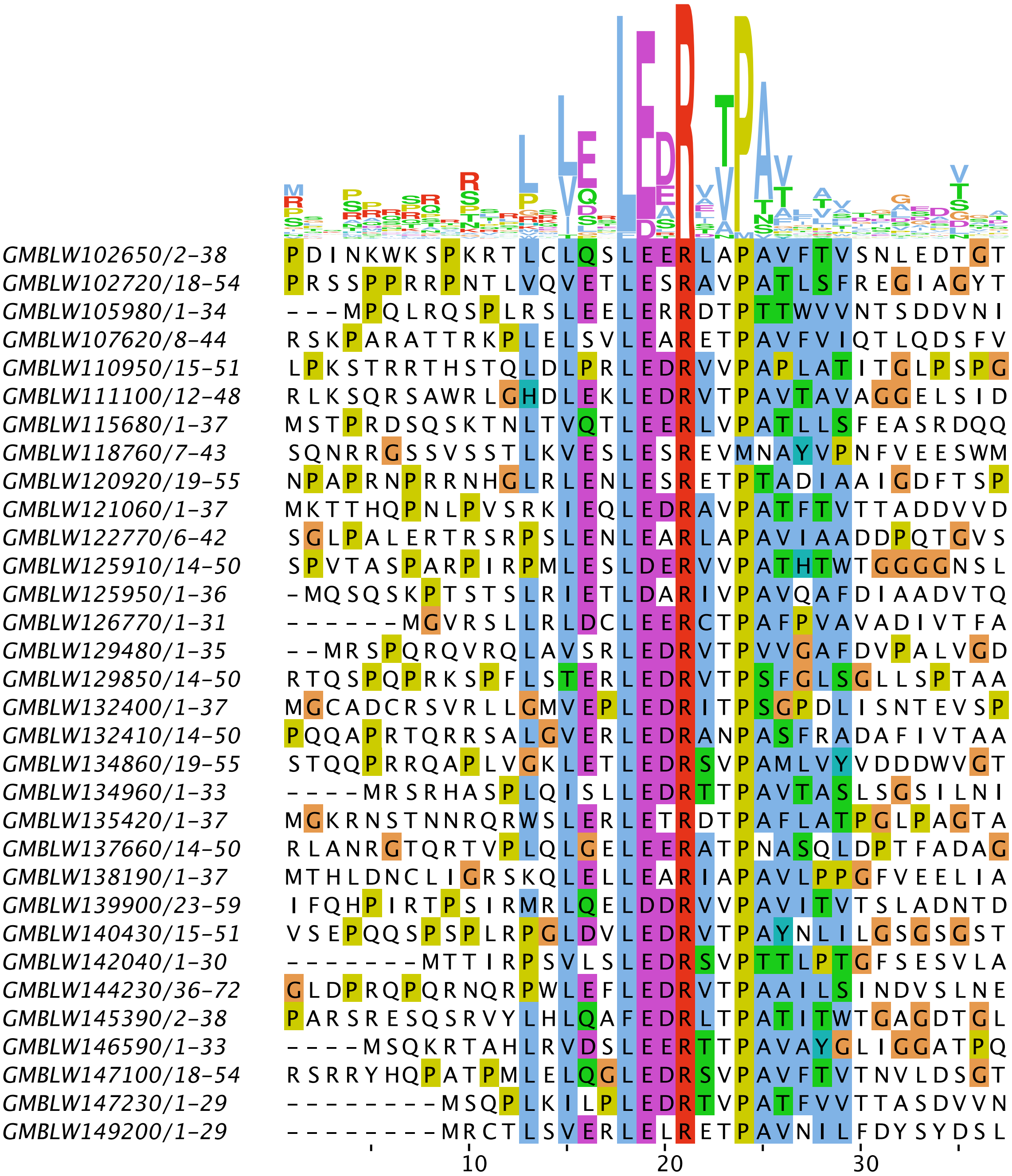


Figure C.1. Genomic map of an operon containing one gene unique to the *Gemmata*/*Tuwongella* clade. Coloured arrows represent genes found in all four species, with blue representing that which is unique to the clade. Arrows with strong colour were identified experimentally in that species. Numbers above arrows show the OrthoMCL cluster ID for that protein. The functional annotations shown above *T. immotidiffusa* are based on InterProScan. Created using genoPlotR.

D. Cell surface signal peptide



```
GMBLW102650/2-38     PDINKWKSPKRTLCLQSLEERLAPAVFTVSNLEDTGT
GMBLW102720/18-54    PRSSPPRRPNTLVQVETLESRAVPATLSFREGIAGYT
GMBLW105980/1-34     ---MPQLRQSPLRSLEELERRDTPTTWVVNTSDDVNI
GMBLW107620/8-44     RSKPARATTRKPLELSVLEARETPAVFVIQTLQDSFV
GMBLW110950/15-51    LPKSTRRTHSTQLDLPRLEDRVVPAPLATITGLPSPG
GMBLW111100/12-48    RLKSQRSAWRLGHDLEKLEDRVTPAVTAVAGGELSID
GMBLW115680/1-37     MSTPRDSQSKTNLTVQTLEERLVPATLLSFEASRDQQ
GMBLW118760/7-43     SQNRRGSSVSSTLKVESLESREVMNAYVPNFVEESWM
GMBLW120920/19-55    NPAPRNPRRNHGLRLENLESRETPTADIAAIGDFTSP
GMBLW121060/1-37     MKTTHQPNLPVSRKIEQLEDRAVPATFTVTTADDVVD
GMBLW122770/6-42     SGLPALERTRSRPSLENLEARLAPAVIAADDPQTGVS
GMBLW125910/14-50    SPVTASPARPIRPMLESLDERVVPATHTWTGGGGNSL
GMBLW125950/1-36     -MQSQSKPTSTSLRIETLDARIVPAVQAFDIAADVTQ
GMBLW126770/1-31     ------MGVRSLLRLDCLEERCTPAFPVAVADIVTFA
GMBLW129480/1-35     --MRSPQRQVRQLAVSRLEDRVTPVVGAFDVPALVGD
GMBLW129850/14-50    RTQSPQPRKSPFLSTERLEDRVTPSFGLSGLLSPTAA
GMBLW132400/1-37     MGCADCRSVRLLGMVEPLEDRITPSGPDLISNTEVSP
GMBLW132410/14-50    PQQAPRTQRRSALGVERLEDRANPASFRADAFIVTAA
GMBLW134860/19-55    STQQPRRQAPLVGKLETLEDRSVPAMLVYVDDDWVGT
GMBLW134960/1-33     ----MRSRHASPLQISLLEDRTTPAVTASLSGSILNI
GMBLW135420/1-37     MGKRNSTNNRQRWSLERLETRDTPAFLATPGLPAGTA
GMBLW137660/14-50    RLANRGTQRTVPLQLGELEERATPNASQLDPTFADAG
GMBLW138190/1-37     MTHLDNCLIGRSKQLELLEARIAPAVLPPGFVEELIA
GMBLW139900/23-59    IFQHPIRTPSIRMRLQELDDRVVPAVITVTSLADNTD
GMBLW140430/15-51    VSEPQQSPSPLRPGLDVLEDRVTPAYNLILGSGSGST
GMBLW142040/1-30     -------MTTIRPSVLSLEDRSVPTTLPTGFSESVLA
GMBLW144230/36-72    GLDPRQPQRNQRPWLEFLEDRVTPAAILSINDVSLNE
GMBLW145390/2-38     PARSRESQSRVYLHLQAFEDRLTPATITWTGAGDTGL
GMBLW146590/1-33     ----MSQKRTAHLRVDSLEERTTPAVAYGLIGGATPQ
GMBLW147100/18-54    RSRRYHQPATPMLELQGLEDRSVPAVFTVTNVLDSGT
GMBLW147230/1-29     --------MSQPLKILPLEDRTVPATFVVTTASDVVN
GMBLW149200/1-29     --------MRCTLSVERLELRETPAVNILFDYSYDSL
```

Figure D.1. Alignment and sequence logo of protein sequences from the experimentally verified *T. immotidiffusa* proteome, with hits to the motif **Lx[VL]ExLEDRx[VT]PA** at most 70 residues from the N-terminus and with up to three mismatches. The sequences were aligned around the identified motif and used to create the sequence logo with the WebLogo server at http://weblogo.berkeley.edu/logo.cgi.

References

[1] N Gimesi. *Hydrobiologiai tanulmanyok*. Kiadja a Magyar Ciszterci Rend, 1924.

[2] C Jenkins and J. T Staley. History, classification and cultivation of the planctomycetes. In J. A Fuerst, editor, *Planctomycetes: Cell Structure, Origins and Biology*, chapter 1, pages 1–38. Springer Science+Business Media, New York, 2013.

[3] M. R Lindsay, R. I Webb, M Strous, M. S Jetten, M. K Butler, R. J Forde, and J. A Fuerst. 2001. Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. *Archives of Microbiology*, 175(6):413–429.

[4] J. A Fuerst and E Sagulenko. 2012. Keys to eukaryality: planctomycetes and ancestral evolution of cellular complexity. *Frontiers in Microbiology*, 3(167):b94.

[5] R Santarella-Mellwig, S Pruggnaller, N Roos, I. W Mattaj, and D. P Devos. 2013. Three-dimensional reconstruction of bacteria with a complex endomembrane system. *PLoS Biology*, 11(5):e1001565.

[6] E König, H Schlesner, and P Hirsch. 1984. Cell wall studies on budding bacteria of the planctomyces/pasteuria group and on a prosthecomicrobium sp. *Archives of Microbiology*, 138(3):200–205.

[7] W Liesack, H Konig, H Schlesner, and P Hirsch. 1986. Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the *Pirella/Planctomyces* group. *Archives of Microbiology*, 145:361–366.

[8] M Strous, E Pelletier, S Mangenot, T Rattei, A Lehner, M. W Taylor, M Horn, H Daims, D Bartol-Mavel, P Wincker, V Barbe, N Fonknechten, D Vallenet, B Segurens, C Schenowitz-Truong, C Médigue, A Collingro, B Snel, B. E Dutilh, H. J. M. O d Camp, C v. d Drift, I Cirpus, K. T v. d Pas-Schoonen, H. R Harhangi, L v Niftrik, M Schmid, J Keltjens, J v. d Vossenberg, B Kartal, H Meier, D Frishman, M. A Huynen, H.-W Mewes, J Weissenbach, M. S. M Jetten, M Wagner, and D. L Paslier. 2005. Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*, 440:790–794.

[9] F. O Glöckner, M Kube, M Bauer, H Teeling, T Lombardot, W Ludwig, D Gade, A Beck, K Borzym, K Heitmann, R Rabus, H Schlesner, R Amann, and R Reinhardt. 2003. Complete genome sequence of the marine planctomycete pirellula sp. strain 1. *Proceedings of the National Academy of Sciences*, 100(14):8298–8303.

[10] J. A Fuerst and E Sagulenko. 2014. Towards understanding the molecular mechanism of the endocytosis-like process in the bacterium gemmata obscuriglobus. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1843(8):1732–1738. Protein trafficking and secretion in bacteria.

[11] D. R Speth, M Van Teeseling, and M Jetten. 2012. Genomic analysis indicates the presence of an asymmetric bilayer outer membrane in planctomycetes and verrucomicrobia. *Frontiers in Microbiology*, 3:304.

[12] H. P Erickson and M Osawa. 2010. Cell division without ftsz–a variety of redundant mechanisms. *Molecular Microbiology*, 78(2):267–270.

[13] T. G. A Lonhienne, E Sagulenko, R. I Webb, K.-C Lee, J Franke, D. P Devos, A Nouwens, B. J Carroll, and J. A Fuerst. 2010. Endocytosis-like protein uptake in the bacterium *Gemmata obscuriglobus*. *Proceedings of the National Academy of Sciences*, 107(29):12883–12888.

[14] A Lieber, A Leis, A Kushmaro, A Minsky, and O Medalia. 2009. Chromatin organization and radio resistance in the bacterium gemmata obscuriglobus. *Journal of Bacteriology*, 191(5):1439–1445.

[15] P. D Franzmann and V. B. D Skerman. 1984. *Gemmata obscuriglobus*, a new genus and species of the budding bacteria. *Antonie van Leeuwenhoek*, 50:261–268.

[16] I. H. M Brümmer, A. D. M Felske, and I Wagner-Döbler. 2004. Diversity and seasonal changes of uncultured planctomycetales in river biofilms. *Applied and Environmental Microbiology*, 70(9):5094–5101.

[17] J Wang, C Jenkins, R. I Webb, and J. A Fuerst. 2002. Isolation of gemmata-like and isosphaera-like planctomycete bacteria from soil and freshwater. *Applied and Environmental Microbiology*, 68(1):417–422.

[18] B. C Berks. 2015. The twin-arginine protein translocation pathway. *Annual Review of Biochemistry*, 84:843–864.

[19] H Mori and K Ito. 2001. The sec protein-translocation pathway. *TRENDS in Microbiology*, 9:494–500.

[20] H Tokuda and S i Matsuyama. 2004. Sorting of lipoproteins to the outer membrane in *E. coli*. *Biochimica et Biophysica Acta*, 5:13.

[21] S. J Facey and A Kuhn. 2010. Biogenesis of bacterial inner-membrane proteins. *Cellular and Molecular Life Sciences*, 67:2343–2362.

[22] E Papanikou, S Karamanou, and A Economou. 2007. Bacterial protein secretion through the translocase nanomachine. *Nature Reviews Microbiology*, 5(11):839–851.

[23] P Chahales and D. G Thanassi. 2015. A more flexible lipoprotein sorting pathway. *Journal of bacteriology*, 197(10):1702–1704.

[24] S Okuda and H Tokuda. 2011. Lipoprotein sorting in bacteria. *Annual review of microbiology*, 65:239–259.

[25] S Fischer, B. P Brunk, F Chen, X Gao, O. S Harb, J. B Iodice, D Shanmugam, D. S Roos, and C. J Stoeckert. 2011. Using orthomcl to assign proteins to orthomcl-db groups or to cluster proteomes into new ortholog groups. *Current protocols in bioinformatics*, pages 6–12.

[26] C. A Schnaitman. 1971. Solubilization of the cytoplasmic membrane of escherichia coli by triton x-100. *Journal of Bacteriology*, 108(1):545–552.

[27] C Filip, G Fletcher, J. L Wulff, and C Earhart. 1973. Solubilization of the cytoplasmic membrane of escherichia coli by the ionic detergent sodium-lauryl sarcosinate. *Journal of bacteriology*, 115(3):717–722.

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

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

[30] L Käll, A Krogh, and E. L Sonnhammer. 2004. A combined transmembrane topology and signal peptide prediction method. *Journal of molecular biology*, 338(5):1027–1036.

[31] L Käll, A Krogh, and E. L Sonnhammer. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the phobius web server. *Nucleic acids research*, 35(suppl 2):W429–W432.

[32] F. S Berven, K Flikka, H. B Jensen, and I Eidhammer. 2004. Bomp: a program to predict integral β-barrel outer membrane proteins encoded within genomes of gram-negative bacteria. *Nucleic acids research*, 32(suppl 2):W394–W399.

[33] W. C Wimley. 2003. The versatile β-barrel membrane protein. *Current opinion in structural biology*, 13(4):404–411.

[34] O Rahman, S. P Cummings, D. J Harrington, and I. C Sutcliffe. 2008. Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of gram-positive bacteria. *World Journal of Microbiology and Biotechnology*, 24:2377–2382.

[35] K Yamaguchi, F Yu, and M Inouye. 1988. A single amino acid determinant of the membrane localization of lipoproteins in e. coli. *Cell*, 53(3):423–432.

[36] J. M Gennity and M Inouye. 1991. The protein sequence responsible for lipoprotein membrane localization in escherichia coli exhibits remarkable specificity. *Journal of Biological Chemistry*, 266(25):16458–16464.

[37] T. N Petersen, S Brunak, G v Heijne, and H Nielsen. 2011. Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–786.

[38] A Krogh, B Larsson, G Von Heijne, and E. L Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580.

[39] J. D Bendtsen, H Nielsen, D Widdick, T Palmer, and S Brunak. 2005. Prediction of twin-arginine signal peptides. *BMC bioinformatics*, 6(1):167.

[40] W Li, A Cowley, M Uludag, T Gur, H McWilliam, S Squizzato, Y. M Park, N Buso, and R Lopez. 2015. The embl-ebi bioinformatics web and programmatic tools framework. *Nucleic acids research*, 43(W1):W580–W584.

[41] S Fuchs. GRAVY Calculator, 2013. `http://gravy-calculator.de` (Accessed June 2016).

[42] J Kyte and R. F Doolittle. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157(1):105–132.

[43] P Jones, D Binns, H.-Y Chang, M Fraser, W Li, C McAnulla, H McWilliam, J Maslen, A Mitchell, G Nuka, et al. 2014. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

[44] R. D Finn, P Coggill, R. Y Eberhardt, S. R Eddy, J Mistry, A. L Mitchell, S. C Potter, M Punta, M Qureshi, A Sangrador-Vegas, et al. 2016. The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–D285.

[45] D Wilson, R Pethica, Y Zhou, C Talbot, C Vogel, M Madera, C Chothia, and J Gough. 2009. Superfamily—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic acids research*, 37(suppl 1):D380–D386.

[46] C. J Sigrist, E De Castro, L Cerutti, B. A Cuche, N Hulo, A Bridge, L Bougueleret, and I Xenarios. 2012. New and continuing developments at prosite. *Nucleic acids research*, page gks1067.

[47] J Bezanson, A Edelman, S Karpinski, and V. B Shah. November 2014. Julia: A fresh approach to numerical computing.

[48] D. J Studholme, J. A Fuerst, and A Bateman. 2004. Novel protein domains and motifs in the marine planctomycete rhodopirellula baltica. *FEMS microbiology letters*, 236(2):333–340.

[49] R Schwartz, C. S Ting, and J King. 2001. Whole proteome pi values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome research*, 11(5):703–709.

[50] E Gasteiger, C Hoogland, A Gattiker, S Duvaud, M. R Wilkins, R. D Appel, and A Bairoch. *Protein identification and analysis tools on the ExPASy server*. Springer, 2005.

[51] M Osborn, J Gander, E Parisi, and J Carson. 1972. Mechanism of assembly of the outer membrane of salmonella typhimurium isolation and characterization of cytoplasmic and outer membrane. *Journal of Biological Chemistry*, 247(12):3962–3972.

[52] B Voigt, C. X Hieu, K Hempel, D Becher, R Schlüter, H Teeling, F. O Glöckner, R Amann, M Hecker, and T Schweder. 2012. Cell surface proteome of the marine planctomycete rhodopirellula baltica. *Proteomics*, 12(11):1781–1791.

[53] S. H Payne, S Bonissone, S Wu, R. N Brown, D. N Ivankov, D Frishman, L Paša-Tolić, R. D Smith, and P. A Pevzner. 2012. Unexpected diversity of signal peptides in prokaryotes. *MBio*, 3(6):e00339–12.

[54] K. V Korotkov, M Sandkvist, and W. G Hol. 2012. The type ii secretion system: biogenesis, molecular architecture and mechanism. *Nature Reviews Microbiology*, 10(5):336–351.