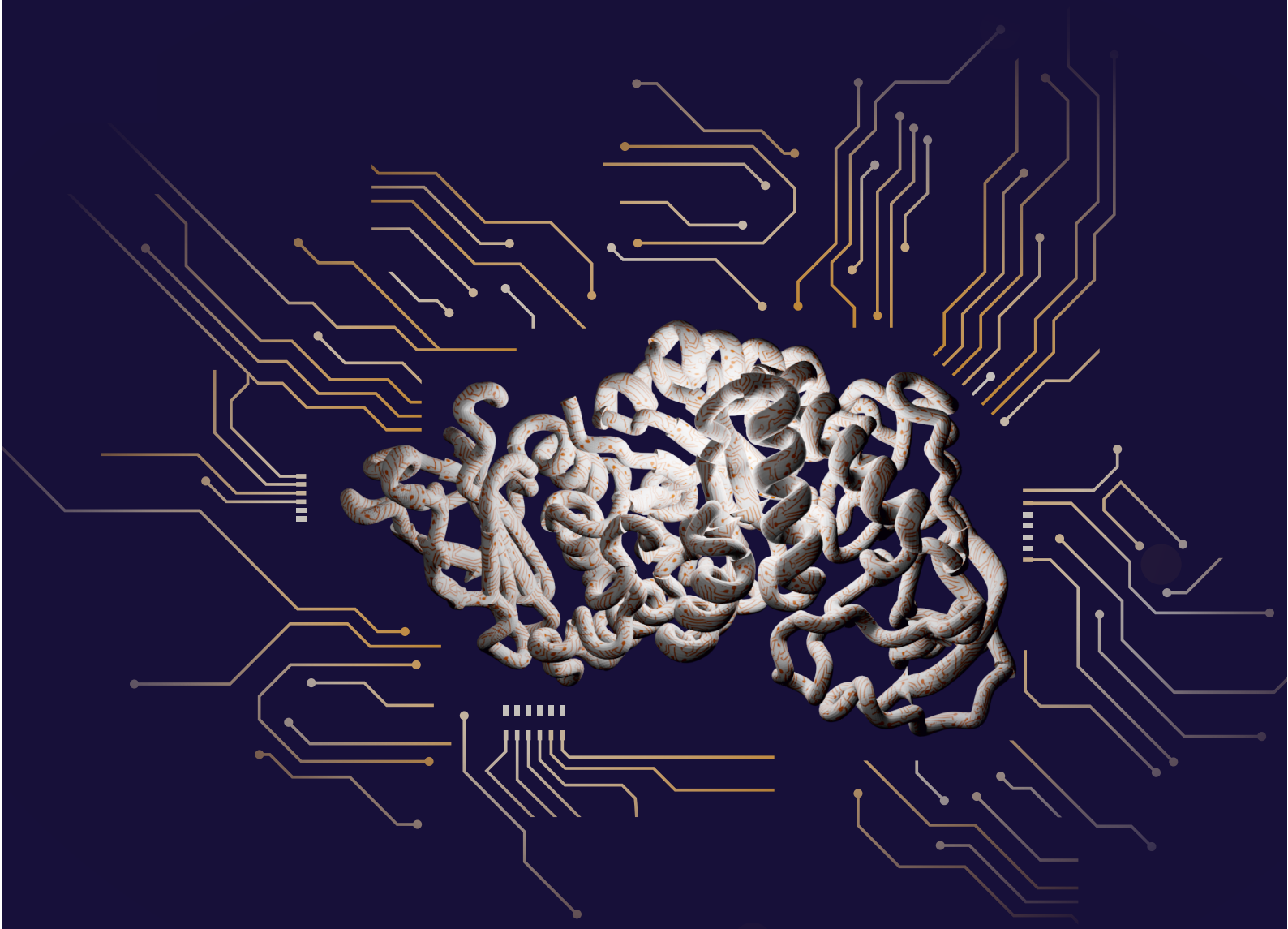
BACTERIAL PROTEIN SORTING: EXPERIMENTAL AND COMPUTATIONAL APPROACHES



STEFANO GRASSO

university of groningen

Bacterial protein sorting: experimental and computational approaches

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of
the Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

16 December 2020 at 18.30 hours

by

Stefano Grasso

born on 27 June 1991
in Vercelli, Italy

Supervisors

Prof. J. M. van Dijl

Dr. T. E. van Rij

Assessment committee

Prof. R.E. Dalbey

Prof. M. Heinemann

Prof. W.J. Quax

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION AND SCOPE OF THE THESIS

Introduction

Biotechnological exploitation of living (micro-)organisms has started some millennia ago[1], probably as a cause or as a consequence of the development of agriculture and the domestication of animals. The earliest biotechnological products were prototypes of bread[2], beer[3], and later wine[4,5], based on the exploitation of the fermentation process typical of yeasts. Similarly, also the first dairy products, such as yogurt and cheese, can be considered amongst the oldest biotechnological commodities, since for their production enzymes, mainly proteases, are needed. Rennet, the enzyme mix sourced from mammalian stomachs to produce cheese, was the first food ingredient to be replaced by biotechnologically produced chymosin, which was approved by the U.S. Food and Drug Administration in 1991[6]. Yet the market started a few years earlier with the production of enzymes for industrial processes, in particular in the textile industry[7].

The global market of industrial enzymes has been growing ever since, and it is expected to reach the size of 7 billion $ in 2021[7]. Due to this fast growth of the market, combined with the needs for more diverse and cheaper enzymes, many efforts to improve the production in terms of quality and efficiency were undertaken in the last two decades[8].

Microorganism are the favourite source and production tool for industrial enzymes due to their availability, easiness of manipulation and fast growth rates[8]. Bacteria are widely used, but yeast and filamentous fungi are widely employed as well. Because of their broad usage and the fundamental research questions they raise, bacterial protein production and sorting pathways have received much attention over many years. In fact, in order to make enzymes more profitable, and to expand the current list of industrial enzymes, biochemical and genetic tools have been employed to both understand and hack how proteins are produced and where they are transported, within or outside the cell[7,8].

In the last decade, a plethora of computational tools was developed with the aim of supporting and improving metabolic and protein engineering, involving a multitude of different approaches[9,10]. Computational tools may be based on very different types of algorithms and may have different specificities, but they all share the goal of reducing the amount of experimental work and, thus, both time and expenses, by *in* silico simulating and predicting *in vivo* processes.

In the present dissertation, the main focus will be placed on Gram-positive bacteria and their protein sorting and secretion pathways, using *Bacillus subtilis* as a model organism. In particular the classical secretion pathway will be addressed, which is commonly used to obtain secreted biotechnological products in order to ease downstream recovery and the subsequent purification of proteinaceous products. Additionally, computational tools related to protein sorting and subcellular protein localization (SCL) will be discussed. A specific discussion will be dedicated to the question how bacterial cells select and determine which proteins are to be secreted and how computational tools can help in reproducing and predicting the related bacterial behaviour.

Protein sorting in Gram-positive bacteria

Before discussing how proteins are targeted to the different cellular compartments, it is necessary to determine and define these compartments. Unfortunately, multiple layers of semantic issues have confounded the terminology that is often used[11]. Gram-positive bacteria were traditionally identified based on the retention of a chemical stain, named the Gram stain after the name of its inventor, in the outer part of the cell envelope, which is now known as the cell wall[12]. In contrast, the outer membrane of Gram-negative bacteria does not retain the crystal violet-iodine complex used for the Gram staining. Upon discovery of the cellular structures of different bacteria, the Gram-staining was linked to the presence of a single or double membrane, dividing bacteria into Gram-positive and Gram-negative species, based on differences in their membrane enclosure. Nevertheless, taxonomically speaking, it has become more and more challenging to separately cluster these

two groups[13]. In the past, when phylogenetic analyses based on genome sequences were not yet possible, the equivalency of the terms 'Gram-positive = Firmicutes (including Actinobacteria) = monoderm', and 'Gram-negative = Gracilicutes = diderm' was widely accepted[14]. Nowadays, with a better understanding of the ultra-structures of bacteria, and improved phylogenetic analyses, this (subdi)vision has become obsolete[13]. Nonetheless, it is widely retained in scientific publications, which may lead to misunderstandings. In the present thesis, the term Gram-positive will be used to refer to monoderm Firmicutes only, as exemplified by the model organism *B. subtilis*, unless stated otherwise. The Firmicutes have a homogenous and well-conserved cellular structure, allowing a generalization of the protein sorting pathways within them.

As graphically represented in Figure 1 Gram-positive bacteria possess a fairly simple cellular envelope, consisting of a plasma membrane and a thick peptidoglycan layer (i.e. the cell wall). Within the cell, proteins can be targeted to four main compartments: the cytoplasm, the plasma membrane, the cell wall and the extracellular milieu. Given that bacterial proteins are ribosomally synthesized in the cytoplasm, all of them start their journey in this compartment and are subsequently targeted or sorted to their final destination. Additionally, in order to reach the extracellular milieu (i.e. the furthest destination from the place of synthesis), proteins need to transiently cross the plasma membrane and the cell wall[11,15,16]. In some cases, more specificity can be given to the localization of proteins, and this will be discussed separately for each SCL. Moreover, it must be noted that proteins may also have multiple or alternative SCLs, which may depend on internal or external stimuli.

From cytoplasm to plasma membrane

As the journey of all proteins starts in the cytoplasm, the journey of most cytosolic proteins will be quite short, essentially ending with their synthesis or assembly into protein complexes. Instead, all other proteins need to be targeted towards the plasma membrane as a first step. This can generally take place either co-translationally or post-translationally. The vast majority of membrane proteins appears to be co-translationally targeted to the membrane, while for secreted and cell wall-associated (CW) proteins both mechanisms may be employed. Additionally, two fundamentally different secretion pathways co-exist in many bacteria: 1) the general secretion (Sec) pathway, which secretes most proteins co- or post-translationally in an unfolded state; and 2) the twin-arginine translocation (Tat) pathway, which secretes proteins exclusively in a folded state and thus only post-translationally[16–20].

At this point a premise is necessary: most of the targeting and translocation mechanisms have only been studied in *Escherichia coli* and are thus relatively well understood in Gram-negative bacteria. As a consequence, various aspects of protein translocation still remain to be clarified in *B. subtilis*, although they may be very similar to their counterparts in *E. coli*. In particular, in Gram-negative bacteria, proteins can be routed to the Sec pathway post-translationally with the aid of SecB, a chaperone that helps maintaining them in an unfolded state. A homologue of SecB is absent from *B. subtilis*, but a possible analogous protein, the chaperone CsaA, has been identified several years ago. Nevertheless, it is still debated to what extent proteins are co- or post-translationally secreted via the Sec pathway in *B. subtilis*[19,21,22].

Once the nascent protein emerges from the ribosomal exit tunnel, it can be recognized by the signal recognition particle (SRP), a riboprotein complex consisting of an RNA molecule of 271 nucleotides named scRNA, the Ffh protein, and the HBsu protein (also known as HupA or Hbs)[21]. SRP is able to recognize a hydrophobic region in the nascent peptide and weakly bind to it. This hydrophobic region can either be the signal peptide (SP) necessary to translocate proteins across the plasma membrane, i.e. to secrete them, or the first hydrophobic α-helix of a trans-membrane (TM) protein[16,19]. Once the nascent chain is bound by SRP, it is co-translationally targeted with the cognate ribosome, to FtsY. FtsY is a peripheral membrane protein that acts as a receptor of SRP and mediates its docking with the Sec translocase[20]. The translocase complex is composed of a motor, the ATPase SecA, which provides the energy for translocation[22]; the SecYEG channel

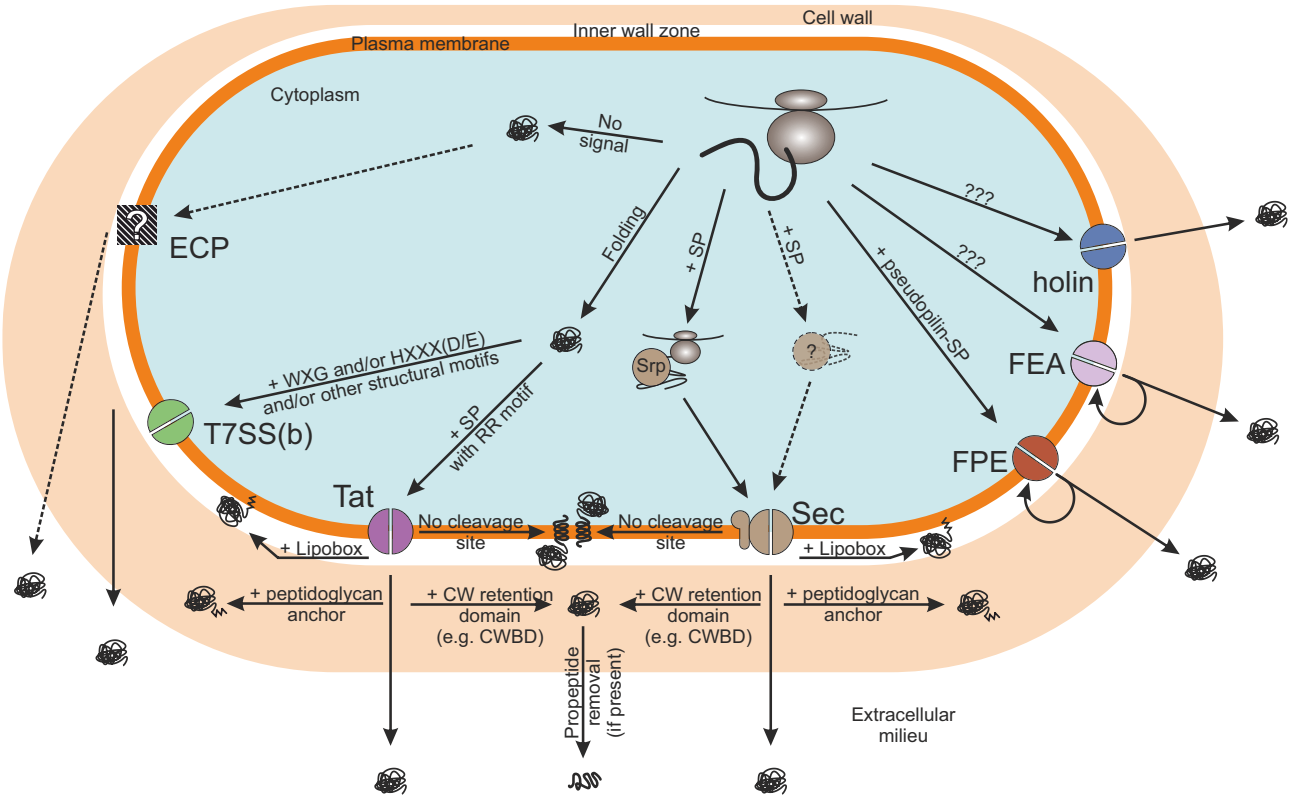


Figure 1. Summary of the known protein secretion pathways in Firmicutes. Cartoon representing a Firmicute bacterium and its different secretion pathways. On the plasma membrane (dark orange) are represented the different translocation apparatuses. In the central position is indicated the general secretory pathway (Sec), which translocates the majority of proteins either co-translationally or post-translationally across the membrane. Proteins are targeted to the Sec translocon via the signal recognition particle (SRP) or other less well-defined chaperones. The sorting signal is the canonical signal peptide (SP). The substrates of Sec are lipoproteins, integral membrane proteins, cell wall proteins and extracellular proteins. To the left of Sec is indicated the twin-arginine translocation (Tat) pathway, which secretes proteins in a folded state, driven by a canonical signal peptide containing a twin-arginine (RR) motif in its N-region. The known substrates of Tat are integral membrane proteins, cell wall proteins and extracellular proteins. Minor pathways for protein secretion, presented clockwise from right to left include: a) holin-mediated secretion for which the sorting signals and mechanisms are not yet properly elucidated; b) the flagellar export apparatus (FEA) that assembles flagella and other motility organelles, and for which the sorting signals are not fully elucidated; c) the fimbrilin-protein exporter (FPE) that is used to secrete and assemble fimbrial, pseudopilus and competence proteins, and for which the sorting signal is the pseudopilin signal peptide. On the left are indicated in clock-wise order: a) the type VII secretion system (T7SS), present mostly in its functional but incomplete T7SSb form, for which the sorting signal seems to be a structural motif that has not been fully elucidated; and b) the non-canonical secretion events that lead to the appearance of extracellular cytoplasmic proteins (ECPs) in the extracellular milieu, and whose secretion rationale, mechanism and signals are presently unknown. The fate of many proteins translocated across the membrane is to become extracellular proteins. However, proteins exported via Sec or Tat may also be: i) retained in the plasma membrane as a lipoprotein through lipidation of the conserved Cys residue in the lipobox; ii) retained in the cell wall either by covalent attachment to the peptidoglycan, or by interaction of specific retention motifs with the cell wall. Importantly, also transmembrane proteins are translocated through the Sec machinery, but instead of being completely translocated, they are laterally sorted from the translocon and in most cases miss a signal peptidase cleavage site. Of note, some proteins possess a Pro-peptide that supports the post-translocational folding process and keeps the enzyme inactive prior and during membrane translocation. The Pros are usually cleaved, and subsequently degraded, once the protein has reached the cell wall or extracellular milieu.

spanning the plasma membrane and being conserved also in Eukaryotes (Sec61)[20]; and the membrane-integrated chaperone SecDF that employs the proton-motive force to enhance protein translocation[23–25].

At this point the journey of integral membrane proteins and those proteins fully translocated across the membrane diverges. After a continuous channel between the ribosomal exit tunnel and the SecYEG translocon has been formed, the polypeptide chains of integral membrane proteins will be completed and laterally inserted into the plasma membrane[20,26]. During this step, also another protein, YidC, may facilitate the lateral insertion and release of proteins into the membrane. The bacterial YidC insertase was first characterized in *E. coli.* In *B. subtilis*, two homologues of YidC are present: SpoIIIJ (YidC1) and YqjG (YidC2). The first, SpoIIIJ, seems to be the main YidC, but it is not essential and it is involved in sporulation. Notably, only a depletion of both proteins is lethal[18,27,28]. Importantly, studies in *E. coli* have shown that YidC can also autonomously insert (i.e. without the involvement of SecYEG) small membrane proteins with only one or two TM helices into the membrane, but the mechanism is still not completely understood[17,27,29].

Membrane crossing: canonical Sec-mediated translocation

All non-cytosolic and non-membrane proteins that are targeted to the plasma membrane by SRP or alternative chaperones are doomed to be translocated to the extracytoplasmic side. Once the nascent or already completed polypeptide chain is docked to the SecYEG translocon, SecA dimers bound to the translocon will take over and start providing the necessary energy to push the soon-to-be-secreted protein through the translocation channel. To do so, SecA will go through the following cycle of reactions: 1) binding of ATP to SecA will occur prior to the binding of SecA to SecYEG; 2) SecA will subsequently hydrolyse the bound ATP molecule, giving freedom of movement to the (nascent) polypeptide chain, thereby allowing a portion of the polypeptide to slide trough the SecYEG channel; 3) at this point SecA can either dissociate from the SecYEG complex and bind a new molecule of ATP to start the cycle over, or use the energy generated by the dissociation of ADP to push further the polypeptide chain[15,19,30,31]. It must be underlined that during step 2) multiple conformational changes occur, both in SecA to release the secretory polypeptide and in SecYEG, where the channel slightly opens to allow the polypeptide chain to slide through. Although the overall process has been studied in great detail, the exact mechanism used by SecA to drive protein secretion is still not completely clear. As reviewed in detail by Collinson[32], two main models of the mechanism employed by SecA have been formulated: a power stroke/diffusional hybrid and a diffusional ratchet model. In both models the energy is provided through SecA-mediated ATP hydrolysis, together with the proton-motive-force. Possibly, also an optimal involvement of chaperones to keep the nascent chain unfolded, combined with protein folding of the secreted peptide on the extracytoplasmic side of the membrane, will help in preventing a backward movement of the translocating polypeptide[32].

It must be remarked that during the early steps of translocation, the SP acquires a “reversed” hairpin-like position in the membrane, with its N-terminus pointing toward the cytosol. Subsequently, the remaining portion of the secreted protein will unloop and pass through the SecYEG channel. Upon unlooping of the SP, its C-region emerges from the plane of the plasma membrane and a type I signal peptidase (SPase) will cleave the SP from the mature protein. This cleavage releases the latter into the extracytoplasmic space, and leaves the SP within the membrane. In *B. subtilis* five chromosomally-encoded SPases are known, namely SipS, SipT, SipU, SipV, and SipW; additionally, a plasmid-encoded SPase, SipP, has been detected[30]. Lastly, in order to avoid accumulation of SPs within the membrane, they are degraded by a signal peptide peptidase (SPPase)[33]. Different membrane-associated proteases have been invoked in the process of signal peptide degradation, in particular SppA and RasP[34]. However, SppA has probably only a minor role in this process, if any[35].

Membrane crossing: Other pathways

The Sec pathway described in the previous section is regarded as the classical or canonical protein secretion

route in Gram-positive bacteria. However, variants of this pathway have been identified in *B. subtilis* and other organisms. In addition, dedicated pathways exist for the export of specific groups of proteins from the cytoplasm, e.g. flagellar proteins.

The following pathways play no or only a marginal role in the context of industrial protein production. Yet, they are important when trying to understand bacterial sorting mechanisms in their completeness and, thus, for protein subcellular localization (SCL) prediction, which will be discussed in the second half of this chapter.

SecA2

Some pathogenic Gram-positive bacteria, both Firmicutes and Actinobacteria, possess two SecA proteins, namely SecA1 and SecA2. While SecA1 exerts the "normal" function of SecA and is, thus, involved in the translocation of most proteins, SecA2 is involved in the secretion of only a subset of proteins, often virulence factors[36,37]. Despite SecA2 being a paralogue of SecA, it is often smaller in size and, interestingly, it is not conserved among Gram-positive bacteria. This suggests that it probably evolved independently among the various species[37,38]. While the necessity of a SecA paralogue is not fully understood, nor its precise mechanism, the existence of at least two functional SecA2 pathways has been demonstrated so far. The first one is called the SecA2-only system, and it is most likely associated to the canonical SecYEG translocon. Conversely, the second SecA2 pathway involves an accessory SecY protein, called SecY2, and functions independently from SecYEG[37,38]. Remarkably, it has been challenging to understand what types of proteins are secreted via the SecA2 pathways, how they are recognized, and why they need a dedicated secretion apparatus. Studies on SecA2 from various genera of Gram-positive bacteria showed a lot of peculiarities[39]. However, if SecA2 evolved independently multiple times, it may well be that it has acquired different functions and mechanisms in different genera.

Lipoproteins and SPase type II

Another variation in the canonical Sec pathway is responsible for the export and lipidation of lipoproteins. These proteins are translocated across the plasma membrane and then covalently attached to it, thus localizing on the extracytoplasmic side of the membrane[15,40].

Firstly, lipoproteins are targeted to the secretion machinery, as described above for other proteins, through the Sec pathway. Once being translocated to the extracytoplasmic side of the membrane, the prolipoprotein diacylglyceryl transferase (Lgt), catalyses the transfer of a diacylglycerol molecule to the conserved Cys in position +1 of the mature protein (i.e. the first amino acid after the cleavage of the SP occurs; see the section on 'Signal Peptide (SP) types'). This anchors the protein to the plasma membrane via the acquired lipidic moiety. Subsequently, a type II SPase, namely LspA in *B. subtilis*, which recognizes a different consensus sequences compared to type I SPases, is responsible for cleaving the SP. In most Firmicutes, the lipoprotein sorting pathway stops here. However, in pathogenic *Staphylococcus* species and Actinobacteria a further step, shared with Gram-negative bacteria, is present. In these species, after the SP cleavage performed by SPase II, a phospholipid/apolipoprotein transacylase, is responsible for also N-acylating the Cys in position +1[41]. In *Staphylococcus aureus*, this N-acylation is catalysed by the LnsAB system and it serves to avoid Tol-like receptor 2-mediated detection of the pathogen. In certain Actinobacteria and Gram-negative bacteria, N-acylation of the +1 Cys residue is a conserved signal for translocating the lipoprotein to the outer membrane, despite the different structures of the respective outer membranes[40,42,43]. It must be remarked here that lipoproteins may also be exported through the Tat pathway (see the next section), both in Gram-negative bacteria and Actinobacteria, suggesting that both the Sec and Tat export pathways can transport lipoprotein precursors that are subsequently lipidated and processed by SPase II. Nonetheless, in Firmicutes no evidence for Tat-secreted lipoproteins has been found so far[42]. Furthermore, many lipoproteins produced

by *B. subtilis* and other Firmicutes may end up in the extracellular space due to secondary proteolytic removal of the N-terminally acylated Cys residue[15,40,44].

Twin-arginine translocation (Tat) pathway

The Tat pathway is considered as the other major protein secretion pathway, which is known to secrete many proteins in Actinobacteria[45]. Nevertheless, in Firmicutes only a few cargo proteins have been identified[46,47]. The Tat pathway secretes proteins that are already fully folded and even complexed or associated with their respective co-factors in the cytosol, in contrast to the canonical Sec pathway whose cargo proteins are folded after translocation[46–48]. Similar to the Sec pathway, the Tat machinery and related mechanisms are highly conserved, and present both in Gram-positive and Gram-negative bacteria, as well as in Archaea and in the thylakoid membranes of chloroplasts. Despite its conservation, the Tat pathway was found to be very variable in terms of accessory components and compatibility, even among closely related organisms[46,47,49].

Tat systems are usually composed of a docking complex and a pore complex. The docking complex is needed to recognize and bind the SP of cargo proteins and it is formed by two proteins, a TatA-like protein which has a single membrane-spanning helix, and a TatC protein which has 6 TM helices. In Gram-negative bacteria, the docking complex is formed by TatB, which belongs to the family of TatA-like proteins, and TatC. On the contrary, *B. subtilis* possesses a minimal set of Tat proteins, encompassing only TatA and TatC, but no TatB. Remarkably though, there are two copies of each component, encoded by the *tatAd tatCd* and *tatAy tatCy* operons. Within this context, both TatAd and TatAy can play the equivalent role of *E. coli* TatB, and form a docking complex with TatCd or TatCy, respectively. Once the docking complex interacted with the SP, the TatC component is responsible for inserting the cargo protein into the plasma membrane. In turn, the just-formed complex, composed of Tat proteins and the cargo, recruits more TatA proteins necessary for the formation of the 'pore complex'. Lastly, the cargo protein is translocated to the extracytoplasmic side of the plasma membrane. To date, the precise details of this step are not yet elucidated, but it either involves the formation of a channel that may vary in size, or local weakening of the membrane[46–48]. It is interesting to note that in *B. subtilis* both the TatAdCd and TatAyCy complexes are sufficiently complete to be fully functional. The main difference found so far between the two complexes lies in their specificity for particular cargo proteins, where the phosphodiesterase PhoD is specifically translocated by TatAdCd, while the Dyp-type peroxidase EfeB (YwbN), the Rieske protein QcrA and the metallophosphoesterase YkuE are specifically translocated by TatAyCy[46–48].

Surprisingly, in *B. subtilis* a third copy of TatA, called TatAc, was detected, but its role is not completely clear as it is not essential for secretion. In fact, despite being able to form complexes with both TatCd and TatCy, it is not able to replace either TatAd or TatAy, but it can still support protein secretion by TatAyCy[50].

Type VII(b) secretion system

Traditionally, secretion systems have been numbered with ordinals only in Gram-negative bacteria, ranging from type I to type VI. Nevertheless, when a novel secretion system was discovered in *Mycobacterium tuberculosis*, which belongs to the Actinobacteria, the assignment of a proper name posed challenging. Initially, based on the name of the main secreted factor, ESAT-6 (early secreted antigen target 6; also called EsxA), the system was named ESX. Eventually, *M. tuberculosis* was shown to possess five of such systems named ESX-1 to ESX-5, which are involved in the secretion of virulence factors. Subsequently, as *M. tuberculosis* is a diderm organism, the novel system was also referred to as the type VII secretion system (T7SS)[51–53].

Although it was initially believed that the T7SS would be restricted to *Mycobacteria*, this system is also active in other Gram-positive bacteria, both Actinobacteria and Firmicutes. Particularly, in Firmicutes the T7SS is present as a simpler variant that does not necessarily seem to be associated with virulence, as exemplified by its presence in non-pathogenic species, such as *B. subtilis*. This minimal T7SS was therefore

named T7SSb. An older name for this secretion system in Firmicutes is WXG100 secretion system (Wss), which was derived from the name of a class of cargo proteins[51–56].

The ESX-1, ESX-3 and ESX-5 systems from *M. tuberculosis* are the best characterized secretion systems of this type. Nevertheless, the roles of many components of this secretion machinery have not been elucidated yet. In Firmicutes, the best studied T7SSb is the one of *S. aureus*, but similar to the situation in *B. subtilis*, some components of this machinery still need to be identified[51,53,56].

Unfortunately, also the cargo proteins of the T7SS are still to be comprehended. It is known that, overall, the systems of this type secrete proteins belonging to many different protein families, including the WXG100, LXG, DUF2563, DUF2580, PE, PPE, Esx and Esp families, all belonging to the EsxAB clan protein superfamily (PFAM: CL0352)[56–58]. These proteins possess some conserved motifs such as the Trp-X-Gly motif, which is present in all T7SS cargo proteins and led to the name WXG100 family, and the Pro-Glu or Pro-Pro-Glu motifs that are specific for the PE and PPE families, respectively. An additional motif, present at the C-terminus of cargo proteins and apparently necessary for their secretion is the H-X-X-X-Asp/Glu-X-X-h-X-X-X-H motif ('H' stands for highly conserved hydrophobic residue, while 'h' for a less conserved one). Of note, in *B. subtilis* only one cargo protein of the T7SS is known, which is secreted as a dimer[54]. This is the YukE protein of unknown function[53].

Non-canonical secretion

Despite the many secretion systems described, there are still secreted proteins that are not translocated across the plasma membrane via one of the aforementioned pathways. In fact, flagellar proteins are exported via a dedicated machinery called the flagellar export apparatus (FEA)[11,15,59], while fimbriae, pseudopili and competence proteins are translocated via a fimbrilin-protein exporter (FPE)[11,30]. An additional secretion system exists for the export of phage-derived proteins, which is named holin, based on its capability to form pores within the bacterial membrane[11,15].

Even taking into account these three additional secretion systems, the membrane translocation of various proteins experimentally detected on the outside of the cell can currently not be attributed to any known secretion system. Most remarkably, many of these proteins have known cytosolic functions. Different hypotheses have been proposed for the mechanisms by which these 'extracellular cytoplasmic proteins' (ECPs) are secreted, and their possible extracellular functions. These range from cell lysis to a still undetected secretion machinery or vesicle-mediated secretion[60–63]. For sure, the most surprising feature of these proteins is the, apparent, lack of a 'secretion' signal that distinguishes them from other cytosolic proteins. In *B. subtilis* it was shown that the amount of non-canonically secreted proteins increases upon the successive deletion of genes for secreted proteases, suggesting that the detectable accumulation of ECPs is directly controlled by proteolysis[64]. In this respect, it is noteworthy that secreted proteases of *B. subtilis* control the activity of autolysins, indicating that the release of ECPs from the cell is to some extent related to lysis[64].

Cell-wall retention signals

Resuming the journey of proteins towards the extracellular milieu, only CW and extracellular proteins are left to be discussed. First of all, it is important to underline that both classes of proteins may be translocated across the plasma membrane in any of the above-described ways. Secondly, all proteins crossing the plasma membrane can eventually become secreted proteins. In essence, some of them are retained within the CW for a certain amount of time and, if no interaction between a particular protein and the CW occurs, it will diffuse into the extracellular milieu becoming an extracellular protein[65,66].

There are two main ways to retain proteins within the CW, namely through covalent attachment to the CW peptidoglycan, or through non-covalent bonds. In the first case, proteins to be attached to the CW, possess both a SP at the N-terminus, necessary to route it out of the cytoplasm, and a conserved motif at

the C-terminus. This conserved C-terminal signal is composed of the consensus sequence Leu-Pro-X-Thr-Gly (LPXTG), a hydrophobic domain and a positively charged domain (the most C-terminal part). This signal is recognized by sortases, a family of transpeptidases, whose role is to cleave the sorting signal between the Thr and Gly residues, and subsequently to covalently bind the Thr residue to the peptidoglycan[15,66,67]. With time it has become clear that multiple sortases are present in Gram-positive bacteria. Sortase A (SrtA) is the most common and representative with the highest number of target proteins, i.e. those with an exact LPXTG consensus sequence. Sortase classes B, C and D are also present among both Firmicutes and Actinobacteria, and they recognize slightly different consensus sequences, such as NP[Q/K]TN, NQPTN, LPXTA, or LAXTG[65–67].

While the covalent attachment of proteins to peptidoglycan is mostly understood as it is a fairly homogeneous process, retention of proteins via non-covalent bonds is definitely less clear and it can be achieved via multiple and different domains. Generally, non-covalently CW-bound proteins possess, within the retention domain, specific motifs that, once folded, can bind to the peptidoglycan or somehow interact with it. Examples of these domains are the cell wall binding domains 1 and 2 (CWBD1 and CWDB2), the lysin motif domain (LysM), the GW module (composed of Gly-Trp dipeptides), the S-layer homology domain (SLHD), the peptidoglycan-binding domain 1 (PBD1), the WXL domain (from the Trp-X-Leu motif), and the clostridial hydrophobic domain (ChW). Each of these binding domains has a specific structure and mode of action. For instance, some domains can simply bind peptidoglycan, while others need to be present in tandem or even in multiple repeats[65,66]. Some domains and motifs have been identified due to their presence in many CW proteins. However, it remains to be proven for some of them that they are actually sufficient for peptidoglycan binding, an example being the SH3b (src Homology-3 bacterial) domain[65].

Other domains are often associated with retention in the CW, for instance the NLPC/P60 domain or the N-acetylmuramoyl-L-alanine amidase domain. However, these are domains with cell wall-modifying enzymatic activities and cannot be considered as proper CW retention signals. In fact, they will bind their CW-derived substrate molecules regardless of their cellular localization or, in case no site to be processed is found, they will not be retained at all[65].

Signal Peptide (SP) types

As already briefly mentioned above, in order to determine the pathway that a protein will follow, a signal must be embedded within its amino acid sequence. Such signals are most often present at the N-terminus of proteins, but a few exceptions exist.

Sec-SPs

Classical Sec-type SPs were the first to be identified, possibly due to their relatively high abundance compared to other sorting signals and their peculiar structure. In fact, they can be virtually divided into three distinct and specific regions: 1) the N-region, approximately 5-6 amino acid residues long, is characterized by the presence of positively charged residues; 2) the H-region, about 15 residues long, is highly hydrophobic and adopts an α-helical structure that will facilitate insertion into the plasma membrane; and 3) the C-region that is usually fairly short (i.e. approximately as long as the N-region or less) and includes a cleavable consensus sequence. Of note, between the end of the H-region and the beginning of the C-region, SPs usually contain a helix-breaking residue. This Pro or Gly residue breaks the α-helix of the H-region, allowing for a less structured C-region that thus becomes accessible to proteases for cleavage[15,30,68].

Despite the common structure of Sec-SPs, they have a little conserved amino acid sequence. This, together with the fact that the secretion efficiency provided by each SP differs, depending on the mature protein it is fused to[69,70], has limited the comprehension of what are the most relevant characteristics of SPs. As extensively reviewed[68], in order to build a comprehensive model of protein secretion, many SP features have been investigated singularly, and even machine learning (ML) models were trained to recognize SPs[71].

Nonetheless, little light has been shed so far on the combined features that determine the SP efficiency. Many characteristics are in fact known to be important, e.g. charge, hydrophobicity, length and the consensus sequence for SPase cleavage, but the respective impact on protein secretion could not be quantified yet. This lack of knowledge and understanding has hampered the *in silico* design of efficient SPs, although a recent advancement was achieved through a ML model able to generate protein-specific SPs, resulting to be functional in 48% of cases[72].

Tat-SPs

Very similar in overall structure to Sec-SPs, Tat-SPs have as their main characteristic the conserved N-terminal S-R-R-x-F-L-K motif, with x being a polar amino acid. This motif, which encompasses two arginine residues has in fact given the name to the whole pathway. Furthermore, the Tat-SPs tend to be slightly longer compared to the Sec-SPs and, despite presenting slightly different statistics for charge and hydrophobicity, they retain the same tripartite structure and consensus sequence for SPase cleavage[15,30,46,47].

It must be remarked that it is difficult to precisely identify Tat-secreted proteins due to two main factors. Firstly, there is a high degree of similarity between the Sec- and Tat-SPs, which makes prediction prone to high false-positive rates. In fact, this similarity can lead a protein to either of the two secretion machineries. Accordingly, Tat-SPs often contain a so-called Sec-avoidance signal in the form of a positively charged residue at the end of the H-region[73]. Secondly, because Tat-secreted proteins are exported almost exclusively in a folded state, it may even happen that they are secreted after the quaternary structure is already formed. Consequently, there may be 'hitchhiker' proteins that are translocated through the Tat translocon, because they are part of a protein complex and not because of a specific sorting signal[46,47].

Lipoprotein-SPs

Similar to Sec- and Tat-SPs, also lipoprotein- (Lipo-) SPs possess a tripartite structure with comparable characteristics. In addition to being shorter than the other two SP types, Lipo-SPs present their main difference in the consensus sequence of the cleavage site, which is recognized by type II SPases. This consensus sequence, known also as the lipobox, corresponds to L-(A/S)-(A/G)-C, where C is the first residue of the mature protein. This Cys residue is strictly conserved as it is needed for lipidation[15,30].

Pseudopilin-SPs

As exemplified by *B. subtilis* competence proteins that assemble into a so-called pseudopilus, the pseudopilin-SPs have a completely different structure compared to Sec-, Tat- and Lipo-SPs, with an N-region that is not specifically charged, followed by a hydrophobic H-region. Interestingly, the pseudopilin SPase cleavage site, with consensus sequence K-G-F, is positioned between the N- and H-regions. Upon processing, the +1 Phe residue is methylated. In *B. subtilis*, the cleavage and methylation are carried out by ComC and the subsequent pseudopilin translocation and assembly follows a dedicated pathway[15,30].

T7SS-SPs

For a long time, the sorting signals of proteins secreted via the T7SS have been investigated, and it was only recently determined that, rather than a sequence motif, the signal is presented as a tertiary structure. T7SS cargo proteins are often secreted as dimers, where each monomer possesses two α-helices separated by the W-X-G consensus sequence, resulting in a helix-turn-helix structure. Interestingly, larger T7SS cargo proteins may be secreted as monomers that form a similar four-helix bundle by themselves. Being common to all T7SS substrates, the bundle of four α-helices seems to be recognized by the secretion machinery, thereby serving as the T7SS sorting signal. Additionally, after the helix-turn-helix structure, and located at the C-terminus, there is the aforementioned conserved H-X-X-X-Asp/Glu-X-X-h-X-X-X-H sequence, which also seems to be

involved in binding the secretion machinery[53,54,74,75].

Pro-peptides and pro-regions

The term Pro-peptide (Pro) designates a region that is often found between the SP and the mature protein, and that is proteolytically removed after secretion. The main role of Pros, which are found usually in enzymes, such as proteases, is to help and catalyse protein folding. Additionally, upon folding, the Pro may keep the enzyme inactive. Such a mechanism is necessary to avoid cellular damage caused, for instance, by extracellular proteases that could become active in the cytosol. Consistent with this idea, it has been shown that certain Pros can work both in *cis* and in *trans*, so both when being associated with the mature protein and when the Pro and mature protein are present as two separate molecules. No general structure for Pros has been identified so far, which is in agreement with the fact that Pros must adopt different structures and conformations for each different class of cognate proteins[76,77].

It has been proposed that Pros may enhance the secretion levels of heterologously produced proteins[78,79]. However, no specific mechanism has been described so far. Most likely, it is not the Pro itself that has an influence on secretion. Instead, it was shown that the first 5 to 15 (and up to 30) residues immediately after the cleavage site (also referred to as pro-region) can modulate, and specifically increase, protein secretion levels[68,80].

Protein sorting-related prediction

The subcellular localization (SCL) of proteins is usually experimentally determined as part of their functional annotation. Unfortunately, however, this is both time-consuming, expensive, and often impractical due to the high number of variables that should be tested. For such reasons, the past three decades have seen an explosion of bioinformatics tools, many of which are devoted to predicting protein characteristics and properties with the amino acid sequence as the sole input.

SP prediction

SP prediction has a long history (extensively reviewed in [81]), with the first manual prediction procedure developed in 1983 and based "simply" on the length of the uncharged region and the maximal hydrophobicity[82]. With time, more complex and accurate methods were developed, first based on features, then on position-weight matrices, and lastly ML algorithms, such as artificial neural networks, hidden Markov models, and support vector machines[81]. Eventually, deep learning methods for SP predictions were developed[83,84]. The most famous tool is SignalP, now at the 5th version, but many other effective tools exist as well, such as PrediSi[85], Phobius[86] or SPEPlip[87], and DeepSig[84].

SP prediction consists of two main parts: the detection of the SP itself, and the determination of the SPase cleavage site (i.e. the N-terminus of the mature protein). While combining the detection of both improves the overall prediction of SPs, different tools may vary considerably in one of the two aspects, e.g. a predictor can be accurate in the cleavage site determination, but not in the SP detection. A third aspect is the distinction of different types of SPs, specifically Sec-SPs, Tat-SPs and Lipo-SPs. As discussed above, these three types of SPs are fairly similar, with only a specific twin-arginine motif in the N-region for Tat-SPs and the lipobox at the cleavage site for Lipo-SPs that distinguish them from Sec-SPs. In order to distinguish SP types, different programs were developed. Examples for the prediction of Tat-SPs are TatP[88], TatFind[89,90] and PRED-TAT[91], while examples for the prediction of Lipo-SPs are LipoP[92], SPEPlip[87] and PRED-LIPO[93]. Of note, SPEPlip is able to discriminate between Sec-SPs and Lipo-SPs. More recently, also SignalP introduced this option. With SignalP 5.0[83] it is now possible to simultaneously discriminate all three types of SPs and their cleavage sites. For the sake of completeness, it should also be mentioned that signatures for Sec-SPs, Tat-SPs and Lipo-SPs exist in all major databases, e.g. Interpro[94], Pfam[57], PROSITE[95] and TIGRfam[96].

Notably, all of the mentioned tools based on ML algorithms were trained on relatively small numbers of SPs (in the order of a few hundreds to a couple of thousands) as derived from *E. coli* and *B. subtilis* for Gram-negative and Gram-positive bacteria respectively, and a few related well-characterized organisms. Consequently, predictions will be biased towards the detection of SPs resembling those in the applied training sets, thereby potentially limiting the detection of SPs from more distantly related organisms or SPs with outlying characteristics. Such a bias, which decreases with the availability of more experimental data, must always be taken into account to avoid self-fulfilling prophecies. In particular, the missed detection of novel ("non-standard") SPs may be caused by a missed prediction which, in turn, limits the possibility to improve SP predictions.

Another important consideration is that the mere presence of a SP is sometimes not enough to determine the fate of a protein, set aside for lipoproteins that always localize to the so-called inner wall zone of Gram-positive bacteria[97], unless they are liberated from the membrane by a secondary cleavage step[15,40]. In fact, after translocation initiated by a Sec-SP or Tat-SP, the protein may be retained in the CW or reach the extracellular milieu. Conversely, the non-detection of a SP does not mean that the protein is not translocated across the plasma membrane, as there are many secretion pathways that are poorly understood and whose cognate sorting signals are not yet recognised. This will be discussed in more detail in the following section.

Other sorting and retention signal predictions

In addition to the many SP predictors, a limited number of other sorting and/or retention signal predictors exists. The latter category includes predictors that are dedicated to the detection of TM helices, which immediately identifies integral membrane proteins. Initially, such predictors were based on the hydrophobicity of consecutive residues, but nowadays all tools employ some sort of learning algorithm. Notable examples of the latter category are TMHMM[98], HMMTOP[99], MEMSAT3[100] and OCTOPUS[101]. Interestingly, in order to improve their results, some tools combine both SP and TM helix predictors. For instance, this is the case with SPOCTOPUS[102], or Phobius[86] and PolyPhobius[103]. While the task of predicting TM helices may seem easier compared to the prediction of SPs, the exact determination of where TM helices start and end is still partially inaccurate, often leading to different numbers of TM helices being predicted for the same protein.

With regard to secretion pathways other than Sec and Tat, the respective sorting signals are often poorly known, limited in numbers, or present only in a single mature protein. This hampers the development of dedicated prediction tools. Few notable exceptions are SecretomeP[60], SecretP[104] and NClassG+[105], of which only SecretomeP was still available at the time of writing of this thesis. The three tools are based on features extracted from sequences of non-classically secreted proteins, which often leads to biased or false results, as exemplified by the inclusion of Sec-secreted proteins in the outputs. Lastly, these tools are not able to determine which possible secretion pathway a protein of interest (POI) follows, but only indicate a higher probability to localise it in the extracellular milieu. Although this is a difficult task, different approaches may drastically improve the prediction of non-classically secreted proteins with precision. As demonstrated by the studies presented in **Chapter 2** of this thesis, starting from the actual knowledge of sorting signals rather than the available experimental data, proves to yield superior SCL predictions. In this regard, it is in fact possible to detect specific protein signatures and motifs, as deposited in online databases, with dedicated scanning tools, such as Interpro[94], Pfam[57], PROSITE[95] and TIGRfam[96]. Only Interpro IDs are listed here, but all of the respective databases in fact possess dedicated entries for the cleavage sites of fimbriae and pilins (IPR012902), T7SS-SPs (IPR010310, IPR041275, IPR000084, IPR006829, IPR000030), non-classical SPs (IPR005877, IPR022263, IPR023833), and a long list of other useful signatures. Despite being often very specific and restricted to a few proteins, or to small classes of proteins, these entries are very helpful, not only in the detection of non-classically secreted proteins, but also in identifying the protein secretion pathway followed by a particular protein.

Similarly, a large number of entries exists for the many known CW retention signals. This involves both the covalently and non-covalently CW-bound proteins, and the respective entries can be exploited to predict CW-associated proteins. Of note, a tool for CW anchor prediction called CW-PRED[106] has been developed, but it is restricted to covalently attached proteins, limiting its applicability.

SCL prediction

In addition to single feature predictors, e.g. for SPs, TM helices or other sorting signals, there is a more comprehensive type of prediction, namely the prediction of the final subcellular compartment where a specific protein is localized. This SCL prediction is a key element within the functional annotation of a protein and it is of interest for multiple reasons, ranging from basic scientific knowledge to medical and industrial applications. For Gram-positive bacteria, there are four main localizations, i.e. the cytosol, plasma membrane, CW, and extracellular milieu. The aforementioned inner wall zone is a somewhat debated compartment that was, so far, never included in any SCL predictor. In this respect, it must be noted that, often, the assignment of a SCL to a protein can actually be a semantic issue[11].

Generally, to assign an SCL to a protein, three main approaches are available (extensively reviewed in [81,107–109]), which are based on the information exploited for the prediction: 1) the physico-chemical properties of the protein; 2) the presence of detectable sorting signals; and 3) the homology with known proteins and the subsequent transfer of an SCL designation. While the first approach is nowadays considered sub-optimal (at least if not combined with other approaches), the signal-based and the homology-based approaches are most extensively exploited. These two latter approaches have their respective advantages and disadvantages as discussed in detail in **Chapter 2** of this thesis. The signal-based approach has some particular benefits, the main one being that it can simulate or reproduce the same sorting signal-based mechanisms that would happen within a bacterial cell. On the other hand, this approach is limited by our current knowledge and understanding of sorting pathways and signals, as well as by the availability of tools for their detection. Such limitations should always be taken into account, though not to penalize the signal-based SCL prediction approach but, rather, to foster research in the direction of detecting and understanding all protein sorting pathways and the respective mechanisms. The knowledge thus generated can subsequently be applied to develop appropriate tools for improved SCL prediction.

Lastly, it has turned out convenient to combine different tools that are either able to detect the same type of signal, e.g. by combining SignalP, Phobius and PrediSi for the detection of SPs, or that will detect different types of signals, e.g. TatP and LipoP. With the first approach a compensation for each tool's possible downsides can be achieved, while the latter approach may lead to a reduction of false SCL predictions. Generally, it has been demonstrated that meta-approaches, based on the exploitation of multiple tools, even of multiple SCL predictors based on different approaches, leads to improved predictions, provided that the weights of individual predictions are assigned properly. This type of approach has been implemented in **Chapter 2** of this thesis, underscoring the superiority of meta-predictors for the genome-wide SCL of proteins.

SP efficiency prediction

One aspect of SP predictions that deserves particular attention, is the prediction of protein secretion efficiency as directed by a specific SP sequence, which was so far not possible. This problem is not merely of scientific interest, but it is also of high industrial relevance, because the costs of production of heterologous recombinant proteins are related to the efficiency of their secretion. In an industrial context, efficient secretion of a protein will be directly mirrored in the respective yields from the fermentation broth. For this reason, much effort has been attributed to understanding and removing various bottlenecks that reduce or hamper protein secretion at the industrial scale, and to understanding which SP sequence best drives the secretion of particular wild-type or recombinant POIs.

At present the most frequently used approach involves the screening of an SP library fused to the POI. This is known to be a very expensive and redundant approach and, to make matters worse, it is necessary to repeat this operation for each individual POI and each production host due to the lack of an adequate theoretical understanding of the underlying principles. While many of the relevant SP features that impact on protein secretion efficiency are known[68], there is no general model available to either predict or explain the resulting secretion level. This is even more so a challenge in the context of different POIs. Despite the fact that the SP has been known for 30 years, only the latest approaches that combine big datasets with ML are able to shine a little bit of light on the features that determine protein secretion efficiency[71], and to achieve predictability of the best possible SP-POI match[72]. Therefore, an approach combining high-throughput screening and ML was implemented in the studies described in **Chapter 5** of this thesis.

Scope of the thesis

In **Chapter 1** of this thesis, a brief introduction on the known bacterial protein secretion pathways and the respective mechanisms is presented, with a special focus on monoderm Gram-positive bacteria. Additionally, the scientific and biotechnological implications of protein secretion, and the relevant bioinformatics prediction tools for protein SCL are discussed.

Chapter 2 presents the GP^4 signal-based meta-predictor to determine protein SCLs in Gram-positive bacteria. Meta-predictors have long been known to outperform individual tools for SP or SCL predictions due to their ability to balance the strengths and weaknesses of their individual constituents. GP^4 represents the first implementation of such a signal-based approach for Firmicutes and Actinobacteria, and it is made available as a simple webserver. Difficulties in the construction of meta-predictors like GP^4 originate from the lack of stand-alone versions of the software components, and lack of standardized and easily programmatically parsable outputs. Additionally, the proposed GP^4 approach leaves space for re-interpretation of the results in the light of an appropriate biological context, e.g. the particular species for which protein SCLs are predicted and its relationship to well-characterized model organisms, or the particular types of analysed proteins (i.e. wild-type or recombinant). A benchmark analysis proved GP^4 to be superior over other widely used SCL prediction tools, both in terms of accuracy of the prediction, as well as quality of the details provided on the sorting pathway(s) accessed by particular proteins.

Chapter 3 reviews the association between *Porphyromonas gingivalis*, a Gram-negative oral pathogen, and rheumatoid arthritis from a molecular perspective. Specifically, either secreted or outer-membrane-bound proteins of *P. gingivalis* seem to play a major role in the etiology of the disease and the generation of auto-antibodies. For this reason, an organism-specific signal-based SCL meta-predictor was developed, particularly taking into account the recently discovered type IX secretion system (T9SS), also called porin secretion system (PorSS), as well as possible non-canonical secretion pathways. By "mapping" all *P. gingivalis* proteins based on their SCL, a better understanding of known, novel, and putative virulence factors was achieved with implications for the identification of potentially druggable targets and inhibitors of virulence.

Chapter 4 of this thesis describes the comparison of community-associated (CA) and hospital-associated (HA) methicillin-resistant *S. aureus* (MRSA) aimed at identifying molecular traits that can be used to separate them. A comparative genomic and proteomic analysis was performed, showing that HA-MRSA and CA-MRSA have two different exproteome profiles. In this regard, a signal-based meta-prediction pipeline was developed to assign SCLs to all proteins detected in the exoproteome. This method was implemented in a way that only proteins with a SP would be classified as secreted, showing extremely high levels of non-classical secretion, either due to unknown, and thus unpredictable secretion mechanisms, or to other unspecific mechanisms.

Intriguingly, proteins predicted to be non-classically secreted had a potentially relevant role in virulence and the epidemiological behaviour of the investigated MRSA isolates.

Chapter 5 presents a completely novel synthetic biology approach to understand and predict the secretion efficiency of a specific SP sequence. By combining high-throughput screening with a ML model and its interpretation, it was possible to perform the first round of a Design-Build-Test-Learn (DBTL) cycle aimed at completely elucidating the main SP characteristics. To this end, a library of approximately 12,000 SPs was screened for their efficiency in directing the secretion of an α-amylase by *B. subtilis*. The resulting data was used to train a Random Forest (RF) model, which in turn was interpreted with the game theory approach SHAP (SHapley Additive exPlanations). Subsequently, the model was used to modify SP sequences in order to obtain α-amylase secretion at a desired level. Additionally, a library of pseudo-randomly designed SPs was *in silico* screened for high performing SPs. Out of the 21 tested pseudo-randomly designed SPs, 7 proved to be equal or superior to the wild-type. This study is of prime importance as it is now possible for the first time to generate an accurate predictive model of SP efficiency. In addition, the study provides important explanations on the general and specific SP characteristics that have a relevant impact on protein secretion efficiency.

The last experimental **Chapter 6** of this thesis illustrates an attempt to elucidate the possible role of Pro-peptides (Pros) in protein secretion, and their application potential. The results show that the contribution of the investigated Pros to protein secretion is, most likely, related to the characteristics of the specific sequences, rather than to their function as Pros. Additionally, by analyzing the growth media of protease-proficient and -deficient strains of *B. subtilis* by mass spectrometry, the proteases involved in the cleavage of specific Pros and the respective cleavage sites were investigated.

Finally, **Chapter 7** summarizes the overall findings outlined in this thesis, and places them within a broader context of the protein secretion field and its industrial applications. Possible improvements of the presented approaches are pointed out so that they may be adopted as standards for future fundamental investigations and the production of high-value proteins.

References

1. Verma, A., Rastogi, S., Agrahari, S. & Singh, A. Biotechnology in the realm of history. *J. Pharm. Bioallied Sci.* **3**, 321 (2011).
2. Arranz-Otaegui, A., Gonzalez Carretero, L., Ramsey, M. N., Fuller, D. Q. & Richter, T. Archaeobotanical evidence reveals the origins of bread 14,400 years ago in northeastern Jordan. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7925–7930 (2018).
3. Liu, L. *et al.* Fermented beverage and food storage in 13,000 y-old stone mortars at Raqefet Cave, Israel: Investigating Natufian ritual feasting. *J. Archaeol. Sci. Reports* **21**, 783–793 (2018).
4. McGovern, P. *et al.* Early Neolithic wine of Georgia in the South Caucasus. *Proc. Natl. Acad. Sci.* **114**, E10309–E10318 (2017).
5. McGovern, P. E. *et al.* Fermented beverages of pre- and proto-historic China. *Proc. Natl. Acad. Sci.* **101**, 17593–17598 (2004).
6. Flamm, E. L. How FDA Approved Chymosin: A Case History. *Nat. Biotechnol.* **9**, 349–351 (1991).
7. Arbige, M. V., Shetty, J. K. & Chotani, G. K. Industrial Enzymology: The Next Chapter. *Trends Biotechnol.* **37**, 1355–1366 (2019).
8. Singh, R., Kumar, M., Mittal, A. & Mehta, P. K. Microbial enzymes: industrial progress in 21st century. *3 Biotech* **6**, 174 (2016).
9. Ebert, M. C. & Pelletier, J. N. Computational tools for enzyme improvement: why everyone can – and should – use them. *Curr. Opin. Chem. Biol.* **37**, 89–96 (2017).
10. Medema, M. H., van Raaphorst, R., Takano, E. & Breitling, R. Computational tools for the synthetic design of biochemical pathways. *Nat. Rev. Microbiol.* **10**, 191–202 (2012).
11. Desvaux, M., Hébraud, M., Talon, R. & Henderson, I. R. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol.* **17**, 139–145 (2009).
12. Coico, R. Gram staining. *Curr. Protoc. Microbiol.* **Appendix 3**, Appendix 3C (2005).
13. Megrian, D., Taib, N., Witwinowski, J., Beloin, C. & Gribaldo, S. One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. *Mol. Microbiol.* **113**, 659–671 (2020).
14. Gibbons, N. E. & Murray, R. G. E. Proposals Concerning the Higher Taxa of Bacteria. *Int. J. Syst. Bacteriol.* **28**, 1–6 (1978).
15. Tjalsma, H. *et al.* Proteomics of Protein Secretion by Bacillus subtilis : Separating the 'Secrets' of the Secretome. *Microbiol. Mol. Biol. Rev. 2* **68**, 207–233 (2004).
16. Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* **6**, 477–487 (2008).
17. Kuhn, A., Koch, H.-G. & Dalbey, R. E. Targeting and Insertion of Membrane Proteins. *EcoSal Plus* **7**, (2017).
18. Kumazaki, K. *et al.* Structural basis of Sec-independent membrane protein insertion by YidC. *Nature* **509**, 516–520 (2014).
19. Harwood, C. R. & Cranenburgh, R. Bacillus protein secretion: an unfolding story. *Trends Microbiol.* **16**, 73–9 (2008).
20. Saraogi, I. & Shan, S. Co-translational protein targeting to the bacterial membrane. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843**, 1433–1441 (2014).
21. Zhang, K., Su, L. & Wu, J. Recent Advances in Recombinant Protein Production by *Bacillus subtilis*. *Annu. Rev. Food Sci. Technol.* **11**, 295–318 (2020).
22. Ling Lin Fu *et al.* Protein secretion pathways in Bacillus subtilis: Implication for optimization of heterologous protein secretion. *Biotechnol. Adv.* **25**, 1–12 (2007).
23. Vörös, A. *et al.* SecDF as Part of the Sec-Translocase Facilitates Efficient Secretion of Bacillus cereus

Toxins and Cell Wall-Associated Proteins. *PLoS One* **9**, e103326 (2014).

24. Bolhuis, A. *et al.* SecDF of Bacillus subtilis , a Molecular Siamese Twin Required for the Efficient Secretion of Proteins. *J. Biol. Chem.* **273**, 21217–21224 (1998).
25. Tjalsma, H., Bron, S. & van Dijl, J. M. Complementary impact of paralogous Oxa1-like proteins of Bacillus subtilis on post-translocational stages in protein secretion. *J. Biol. Chem.* **278**, 15622–32 (2003).
26. Zweers, J. C. *et al.* Towards the development of Bacillus subtilis as a cell factory for membrane proteins and protein complexes. *Microb. Cell Fact.* **7**, 10 (2008).
27. Dalbey, R. E., Kuhn, A., Zhu, L. & Kiefer, D. The membrane insertase YidC. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843**, 1489–1496 (2014).
28. Tjalsma, H., Bron, S. & Van Dijl, J. M. Complementary impact of paralogous Oxa1-like proteins of Bacillus subtilis on post-translocational stages in protein secretion. *J. Biol. Chem.* **278**, 15622–15632 (2003).
29. He, H., Kuhn, A. & Dalbey, R. E. Tracking the Stepwise Movement of a Membrane-inserting Protein In Vivo. *J. Mol. Biol.* **432**, 484–496 (2020).
30. Tjalsma, H., Bolhuis, A., Jongbloed, J. D., Bron, S. & van Dijl, J. M. Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiol. Mol. Biol. Rev.* **64**, 515–47 (2000).
31. de Keyzer, J., van der Does, C. & Driessen, A. J. M. The bacterial translocase: a dynamic protein channel complex. *Cell. Mol. Life Sci.* **60**, 2034–2052 (2003).
32. Collinson, I. The Dynamic ATP-Driven Mechanism of Bacterial Protein Translocation and the Critical Role of Phospholipids. *Front. Microbiol.* **10**, 1217 (2019).
33. Traag, B. A., Pugliese, A., Setlow, B., Setlow, P. & Losick, R. A conserved ClpP-like protease involved in spore outgrowth in Bacillus subtilis. *Mol. Microbiol.* **90**, 160–6 (2013).
34. Saito, A. *et al.* Post-liberation cleavage of signal peptides is catalyzed by the site-2 protease (S2P) in bacteria. *Proc. Natl. Acad. Sci.* **108**, 13740–13745 (2011).
35. Henriques, G. *et al.* SppI Forms a Membrane Protein Complex with SppA and Inhibits Its Protease Activity in Bacillus subtilis. *mSphere* **5**, e00724-20 (2020).
36. Feltcher, M. E. & Braunstein, M. Emerging themes in SecA2-mediated protein export. *Nat. Rev. Microbiol.* **10**, 779–789 (2012).
37. Prabudiansyah, I. & Driessen, A. J. M. The canonical and accessory sec system of gram-positive bacteria. in *Current Topics in Microbiology and Immunology* vol. 404 45–67 (Springer Verlag, 2017).
38. Braunstein, M., Bensing, B. A. & Sullam, P. M. The Two Distinct Types of SecA2-Dependent Export Systems. in *Protein Secretion in Bacteria* 29–41 (ASM Press, 2019).
39. Bensing, B. A., Seepersaud, R., Yen, Y. T. & Sullam, P. M. Selective transport by SecA2: An expanding family of customized motor proteins. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843**, 1674–1686 (2014).
40. Tjalsma, H. & Van Dijl, J. M. Proteomics-based consensus prediction of protein retention in a bacterial membrane. *Proteomics* **5**, 4472–4482 (2005).
41. Gardiner, J. H. *et al.* Lipoprotein N -Acylation in Staphylococcus aureus Is Catalyzed by a Two-Component Acyl Transferase System. *MBio* **11**, 1–18 (2020).
42. Okuda, S. & Tokuda, H. Lipoprotein sorting in bacteria. *Annu. Rev. Microbiol.* **65**, 239–259 (2011).
43. Hutchings, M. I., Palmer, T., Harrington, D. J. & Sutcliffe, I. C. Lipoprotein biogenesis in Gram-positive bacteria: knowing when to hold 'em, knowing when to fold 'em. *Trends Microbiol.* **17**, 13–21 (2009).
44. Sibbald, M. J. J. B. *et al.* Mapping the Pathways to StaphylococcalPathogenesis by Comparative Secretomics. *Microbiol. Mol. Biol. Rev.* **70**, 755–788 (2006).
45. Widdick, D. A. *et al.* The twin-arginine translocation pathway is a major route of protein export in Streptomyces coelicolor. *Proc. Natl. Acad. Sci.* **103**, 17927–17932 (2006).

46. Goosens, V. J., Monteferrante, C. G. & Van Dijl, J. M. The Tat system of Gram-positive bacteria. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843**, 1698–1706 (2014).
47. Goosens, V. J. & van Dijl, J. M. Twin-Arginine Protein Translocation. in *Current Topics in Microbiology and Immunology* vol. 404 69–94 (Springer Verlag, 2016).
48. Frain, K. M., Robinson, C. & van Dijl, J. M. Transport of Folded Proteins by the Tat System. *Protein J.* **38**, 377–388 (2019).
49. Bernal-Cabas, M. *et al.* Functional association of the stress-responsive LiaH protein and the minimal TatAyCy protein translocase in Bacillus subtilis. *Biochim. Biophys. Acta - Mol. Cell Res.* **1867**, (2020).
50. Goosens, V. J., De-San-Eustaquio-Campillo, A., Carballido-López, R. & van Dijl, J. M. A Tat ménage à trois — The role of Bacillus subtilis TatAc in twin-arginine protein translocation. *Biochim. Biophys. Acta - Mol. Cell Res.* **1853**, 2745–2753 (2015).
51. Ates, L. S., Houben, E. N. G. & Bitter, W. Type VII Secretion: A Highly Versatile Secretion System. *Microbiol. Spectr.* **4**, 1–21 (2016).
52. Desvaux, M., Hébraud, M., Talon, R. & Henderson, I. R. Outer membrane translocation: numerical protein secretion nomenclature in question in mycobacteria. *Trends Microbiol.* **17**, 338–40 (2009).
53. Unnikrishnan, M., Constantinidou, C., Palmer, T. & Pallen, M. J. The Enigmatic Esx Proteins: Looking Beyond Mycobacteria. *Trends Microbiol.* **25**, 192–204 (2017).
54. Sysoeva, T. A., Zepeda-Rivera, M. A., Huppert, L. A. & Burton, B. M. Dimer recognition and secretion by the ESX secretion system in Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 7653–7658 (2014).
55. Pallen, M. J. The ESAT-6/WXG100 superfamily – and a new Gram-positive secretion system? *Trends Microbiol.* **10**, 209–212 (2002).
56. Sutcliffe, I. C. New insights into the distribution of WXG100 protein secretion systems. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* **99**, 127–131 (2011).
57. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
58. Poulsen, C., Panjikar, S., Holton, S. J., Wilmanns, M. & Song, Y. H. WXG100 protein superfamily consists of three subfamilies and exhibits an α-helical C-terminal conserved residue pattern. *PLoS One* **9**, e89313 (2014).
59. Calvo, R. A. & Kearns, D. B. FlgM Is secreted by the flagellar export apparatus in Bacillus subtilis. *J. Bacteriol.* **197**, 81–91 (2015).
60. Wang, G. *et al.* How are the Non-classically Secreted Bacterial Proteins Released into the Extracellular Milieu? *Curr. Microbiol.* **67**, 688–695 (2013).
61. Kang, Q. & Zhang, D. Principle and potential applications of the non-classical protein secretory pathway in bacteria. *Appl. Microbiol. Biotechnol.* **104**, 953–965 (2020).
62. Zhao, X. *et al.* Exoproteome Heterogeneity among Closely Related Staphylococcus aureus t437 Isolates and Possible Implications for Virulence. *J. Proteome Res.* **18**, 2859–2874 (2019).
63. Ebner, P. & Götz, F. Bacterial Excretion of Cytoplasmic Proteins (ECP): Occurrence, Mechanism, and Function. *Trends Microbiol.* **27**, 176–187 (2019).
64. Krishnappa, L. *et al.* Extracytoplasmic Proteases Determining the Cleavage and Release of Secreted Proteins, Lipoproteins, and Membrane Proteins in Bacillus subtilis. *J. Proteome Res.* **12**, 4101–4110 (2013).
65. Desvaux, M. & Hébraud, M. Analysis of cell envelope proteins. in *Handbook of Listeria Monocytogenes* 359–393 (CRC Press, 2008).
66. Desvaux, M., Dumas, E., Chafsey, I. & HÃ©braud, M. Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure. *FEMS Microbiol. Lett.* **256**, 1–15 (2006).

67. Siegel, S. D., Reardon, M. E. & Ton-That, H. Anchoring of LPXTG-Like Proteins to the Gram-Positive Cell Wall Envelope. in *Current Topics in Microbiology and Immunology* vol. 404 159–175 (Springer Verlag, 2016).
68. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.* **97**, 422–441 (2018).
69. Brockmeier, U. *et al.* Systematic Screening of All Signal Peptides from Bacillus subtilis: A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-positive Bacteria. *J. Mol. Biol.* **362**, 393–402 (2006).
70. Degering, C. *et al.* Optimization of Protease Secretion in Bacillus subtilis and Bacillus licheniformis by Screening of Homologous and Heterologous Signal Peptides. *Appl. Environ. Microbiol.* **76**, 6370–6376 (2010).
71. Peng, C. *et al.* Factors Influencing Recombinant Protein Secretion Efficiency in Gram-Positive Bacteria: Signal Peptide and Beyond. *Front. Bioeng. Biotechnol.* **7**, 139 (2019).
72. Wu, Z. *et al.* Signal Peptides Generated by Attention-Based Neural Networks. *ACS Synth. Biol.* **9**, 2154–2161 (2020).
73. Bogsch, E., Brink, S. & Robinson, C. Pathway specificity for a ΔpH-dependent precursor thylakoid lumen protein is governed by a 'Sec-avoidance' motif in the transfer peptide and a 'Sec-incompatible' mature protein. *EMBO J.* **16**, 3851–3859 (1997).
74. Houben, E. N. G., Korotkov, K. V. & Bitter, W. Take five - Type VII secretion systems of Mycobacteria. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843**, 1707–1716 (2014).
75. Daleke, M. H. *et al.* General secretion signal for the mycobacterial type VII secretion pathway. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 11342–11347 (2012).
76. Shinde, U. & Inouye, M. Propeptide-mediated folding in subtilisin: The intramolecular chaperone concept. *Adv. Exp. Med. Biol.* **379**, 147–154 (1996).
77. Demidyuk, I. V., Shubin, A. V., Gasanov, E. V. & Kostrov, S. V. Propeptides as modulators of functional activity of proteases. *Biomol. Concepts* **1**, 305–322 (2010).
78. Kakeshita, H., Kageyama, Y., Ara, K., Ozaki, K. & Nakamura, K. Propeptide of Bacillus subtilis amylase enhances extracellular production of human interferon-α in Bacillus subtilis. *Appl. Microbiol. Biotechnol.* **89**, 1509–17 (2011).
79. Kouwen, T. R. H. M. *et al.* Contributions of the Pre- And Pro-Regions of a Staphylococcus hyicus Lipase to Secretion of a Heterologous Protein by Bacillus subtilis. *Appl. Environ. Microbiol.* **76**, 659–669 (2010).
80. Tian, P. & Bernstein, H. D. Identification of a post-targeting step required for efficient cotranslational translocation of proteins across the Escherichia coli inner membrane. *J. Biol. Chem.* **284**, 11396–11404 (2009).
81. Nielsen, H., Tsirigos, K. D., Brunak, S. & von Heijne, G. A Brief History of Protein Sorting Prediction. *Protein J.* **38**, 200–216 (2019).
82. McGeoch, D. J. On the predictive recognition of signal peptide sequences. *Virus Res.* **3**, 271–286 (1985).
83. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
84. Savojardo, C., Martelli, P. L., Fariselli, P. & Casadio, R. DeepSig: Deep learning improves signal peptide detection in proteins. *Bioinformatics* **34**, 1690–1696 (2018).
85. Hiller, K., Grote, A., Scheer, M., Münch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **32**, W375-9 (2004).
86. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* **338**, 1027–1036 (2004).

87. Fariselli, P., Finocchiaro, G. & Casadio, R. SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics* **19**, 2498–2499 (2003).
88. Bendtsen, J. D., Nielsen, H., Widdick, D., Palmer, T. & Brunak, S. Prediction of twin-arginine signal peptides. *BMC Bioinformatics* **6**, 167 (2005).
89. Rose, R. W., Brüser, T., Kissinger, J. C. & Pohlschröder, M. Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. *Mol. Microbiol.* **45**, 943–950 (2002).
90. Dilks, K., Rose, R. W., Hartmann, E. & Pohlschröder, M. Prokaryotic utilization of the twin-arginine translocation pathway: A genomic survey. *J. Bacteriol.* **185**, 1478–1483 (2003).
91. Bagos, P. G., Nikolaou, E. P., Liakopoulos, T. D. & Tsirigos, K. D. Combined prediction of Tat and Sec signal peptides with hidden Markov models. *Bioinformatics* **26**, 2811–2817 (2010).
92. Juncker, A. S. *et al.* Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci.* **12**, 1652–1662 (2003).
93. Bagos, P. G., Tsirigos, K. D., Liakopoulos, T. D. & Hamodrakas, S. J. Prediction of lipoprotein signal peptides in Gram-positive bacteria with a Hidden Markov Model. *J. Proteome Res.* **7**, 5082–5093 (2008).
94. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
95. Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344-7 (2013).
96. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29**, 41–3 (2001).
97. Zuber, B. B. *et al.* Granular Layer in the Periplasmic Space of Gram-Positive Bacteria and Fine Structures of Enterococcus gallinarum and Streptococcus gordonii Septa Revealed by Cryo-Electron Microscopy of Vitreous Sections. *J. Bacteriol.* **188**, 6652–6660 (2006).
98. Krogh, a, Larsson, B., von Heijne, G. & Sonnhammer, E. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
99. Tusnády, G. E. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849–850 (2001).
100. Jones, D. T. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* **23**, 538–544 (2007).
101. Viklund, H. & Elofsson, A. OCTOPUS: Improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* **24**, 1662–1668 (2008).
102. Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. SPOCTOPUS: A combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **24**, 2928–2929 (2008).
103. Käll, L., Krogh, A. & Sonnhammer, E. L. L. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* **21 Suppl 1**, i251-7 (2005).
104. Yu, L. *et al.* SecretP: Identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. *J. Theor. Biol.* **267**, 1–6 (2010).
105. Restrepo-Montoya, D., Pino, C., Nino, L. F., Patarroyo, M. E. & Patarroyo, M. A. NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins. *BMC Bioinformatics* **12**, 21 (2011).
106. Litou, Z. I., Bagos, P. G., Tsirigos, K. D., Liakopoulos, T. D. & Hamodrakas, S. J. Prediction of cell wall sorting signals in gram-positive bacteria with a hidden markov model: Application to complete genomes. *J. Bioinform. Comput. Biol.* **6**, 387–401 (2008).
107. Nielsen, H. Protein Sorting Prediction. in *Methods in Molecular Biology* vol. 1615 23–57 (Humana Press Inc., 2017).

108. Nielsen, H. Predicting Subcellular Localization of Proteins by Bioinformatic Algorithms. in *Current Topics in Microbiology and Immunology* vol. 404 129–158 (Springer Verlag, 2015).

109. Wan, S. & Mak, M.-W. *Machine Learning for Protein Subcellular Localization Prediction*. (DE GRUYTER, 2015).

CHAPTER 2

GP^4: AN INTEGRATED GRAM-POSITIVE PROTEIN PREDICTION PIPELINE FOR SUBCELLULAR LOCALIZATION MIMICKING BACTERIAL SORTING

Stefano Grasso, Tjeerd van Rij, Jan Maarten van Dijl

In press for
Briefings in bioinformatics

Abstract

Subcellular localization is a critical aspect of protein function and the potential application of proteins either as drugs or drug targets, or in industrial and domestic applications. However, the experimental determination of protein localization is time-consuming and expensive. Therefore, various localization predictors have been developed for particular groups of species. Intriguingly, despite their major representation amongst biotechnological cell factories and pathogens, a meta-predictor based on sorting signals and specific for Gram-positive bacteria was still lacking. Here we present GP^4, a protein subcellular localization meta-predictor mainly for Firmicutes, but also Actinobacteria, based on the combination of multiple tools, each specific for different sorting signals and compartments. Novelty elements include: improved cell-wall protein prediction, including differentiation of the type of interaction, prediction of non-canonical secretion pathway target proteins, separate prediction of lipoproteins, and better user experience in terms of parsability and interpretability of the results. GP^4 aims at mimicking protein sorting as it would happen in a bacterial cell. As GP^4 is not homology-based it has a broad applicability and does not depend on annotated databases with homologous proteins. Non-canonical usage may include: little studied or novel species, synthetic and engineered organisms, and even re-use of the prediction data to develop custom prediction algorithms. Our benchmark analysis highlights the improved performance of GP^4 compared to other widely-used subcellular protein localization predictors. A webserver running GP^4 is available at: http://gp4.hpc.rug.nl/

Keywords: Protein subcellular localization prediction, Prediction methods, Homology-based prediction, Sorting signals, Gram-positive, GP^4

Supplementary files avaialble at: https://github.com/grassoste/Thesis-supplementary-files

Background

Subcellular localization (SCL) is a key element in the functional annotation of proteins, their use in biotechnology, and their potential as drug candidates or targets. Ideally, SCL should be determined experimentally. Unfortunately, however, this is time-consuming, expensive, and impractical due to the recent explosion in the numbers of whole-genome-sequenced organisms. For such reasons, multiple approaches to predict SCLs have been developed (extensively reviewed in [1–5]).

Given that the prediction of SCL always starts from the amino acidic sequence of a protein, and the desired output is a designated cellular compartment or the extracellular milieu, the presently available approaches can be categorized based on the method of SCL assignment: 1) physico-chemical properties of the protein; 2) detectable sorting signals; and 3) homology and transfer of knowledge. Each approach has its own advantages and disadvantages but, additionally, there can still be different methods implemented within each category that have their own specific pros and cons [1–5]. In this paper, we address the most relevant aspects that should be taken into account, and present a new protein subcellular localization meta-predictor for Firmicutes, named GP⁴, which is also suitable for Actinobacteria.

Historically, the physico-chemical properties of a protein were the first parameters employed to predict signal peptides (SPs) for protein export from the cytoplasm and SCLs. However, physico-chemical properties by themselves are nowadays considered too broad for obtaining accurate results. Instead, two other approaches are regarded as more promising. SCL prediction based on known sorting signals is probably the most suitable approach, as the detection of specific localization tags embedded within the amino acidic sequence is also what cells do to sort their proteins [5–7]. However, a sufficiently detailed understanding of protein sorting mechanisms in the organism of interest is necessary to identify these localization tags with bioinformatic tools for SCL prediction. On the other hand, homology-based methods infer SCL by transferring the annotation of the best hit of a BLAST search to the query protein [1–4]. This last method is frequently used to functionally annotate genomes, genes and proteins. Unfortunately, however, it was estimated that homology-based annotations in the Gene Ontology (GO) database as of March 2006 showed an error rate of 49%. In contrast, homology-independent methods resulted in estimated error rates between 13% and 18% [8]. Altogether, the combination of a low number [9,10] and biased distribution [11,12] of studied and annotated entries in protein databases has resulted in the percolation of erroneous annotations [13,14]. Moreover, while the transfer of annotations may appear effective [15,16], different similarity thresholds can heavily influence the outcome, and will lead to annotation errors in case of low similarity [10,15,17,18].

In addition to the three afore-mentioned methods for SCL assignment, also hybrid methods have been developed, which exploit the strengths and compensate for the weaknesses of the combined approaches and algorithms. This hybrid category encompasses the most frequently used and reliable SCL predictors, such as PsortB [19], CELLO [20], pLoc-mGpos [21], or Protein Analyst [22].

Given the apparent lack of rational design in protein function or structure, it is important to consider the easiness by which evolution re-uses sequences for novel scopes, nullifying the ‘from sequence to structure to function’, and thus localization hypothesis [10,17,23]. Consequently, only annotations whose primary information source is experimental should be regarded with a certain confidence. Other types of annotation should be considered with care [24], and have in extreme cases led to the propagation of mistakes [25,26]. Yet, experimental verification of protein SCL is also not flawless, as there is always cross-contamination during cell disruption, and it is hard to separate living cells from dead cells and their debris that has been released into the extracellular milieu [27].

Due to the major differences in the cellular structures encountered among the three main kingdoms of life (Bacteria, Eukarya and Archaea), bioinformatics tools generally specialize in SCL predictions for one of these domains of life. Unfortunately, within the Bacterial kingdom, the most common subdivision used is

between Gram-positive and Gram-negative bacteria. This distinction is based on the outcome of Gram-staining with crystal violet rather than the cellular architecture and, consequently, leaves space for misinterpretations [28,29]. Given the different morphology, the possible SCLs to be predicted differ substantially. In Gram-positive bacteria as traditionally defined, there are four classical sub-cellular compartments, namely the cytoplasm, the plasma membrane, the cell wall and the extracellular space. A further fifth compartment has been named the inner wall zone [30], which includes the 'periplasmic' area between the plasma membrane and the cell wall. However, to date the inner wall zone has not been considered by SCL prediction tools. Despite an overall agreement on the different SCLs, there is little consensus amongst the different prediction tools about which proteins should be included in each compartment and the respective terminology [28]. Some proteins are unequivocal regarding their SCL, both from the computational and experimental points of view. Other proteins, pose challenges since they may be experimentally found in multiple compartments, or may have been identified in SCLs that contrast with their in silico predicted SCLs. Additionally, some compartments can either be further subdivided or may be 'atypical', as exemplified by fimbriae, pili or spores, which do in fact possess their own peculiar subdivision (e.g. basal body, spore coat, cortex and core).

A crucial aspect in SCL prediction is its scope, or the origin of the query sequence. This can relate to a wild-type protein from a known or novel organism, or to a synthetically designed protein. Although this issue has been theoretically addressed [31], the latter category has never been thoroughly investigated. This may relate to the, thus far, limited needs to predict SCLs for synthetic proteins, but the design and realization of synthetic organisms is becoming more common and will probably increase in the future [32]. To properly address this kind of synthetic proteins, it is important to notice how they may be decontextualized from their original source or environment, i.e. the original organism. In such cases, it may be misleading to directly assign the SCL retrieved from its closest wild-type homologue to the query sequence.

Lastly, a key aspect demanded by all users, is the ease of interpretation of SCL predictions [7,31]. Here, one also needs to consider context information that cannot be submitted with the query sequence, e.g. the investigated species and its peculiarities or the applied design. One option to solve this dilemma is to increase the customizability and flexibility of the prediction tool, thereby allowing the user to include tailored options.

Taking into account so many aspects of SCL prediction is challenging, and multiple solutions with different pros and cons are possible. Here we present a basic prediction pipeline for Gram-positive organisms called GP^4 (Gram-Positive Protein Prediction Pipeline). In brief, GP^4 is based on already available tools for different aspects of SCL prediction, which mainly rely on sorting signals or motif detection. GP^4 minimizes the usage of homology to avoid the afore-mentioned biases. The GP^4 pipeline is particularly suitable for Firmicutes and it is also effective for Actinobacteria, although it cannot predict their outer membrane proteins. GP^4 is available as a webserver with an easy and user-friendly interface at http://gp4.hpc.rug.nl/ (Figure 1). The only required input is a list of fasta amino acid sequences, that can also be submitted as file. Additionally, GP^4 can be used as a standalone program, but only as a pipeline script to produce the relevant data with the implemented tools, to combine them, and to return the final SCL prediction.

Material & Methods

Rationale and general approach

GP^4 assigns up to 5 SCLs, including the 4 'classical' ones, namely the cytoplasm, trans-membrane (TM), cell wall and extracellular. In addition, GP^4 can also return lipoproteins as a result. Despite the latter not being a proper SCL, it was included in GP^4 as a lipidic retention signal is often more informative with regard to protein sorting than the actual SCL. In contrast to other prediction tools, only integral membrane proteins with one or more TM α-helices are predicted by GP^4 as membrane proteins. In contrast, peripheral membrane proteins

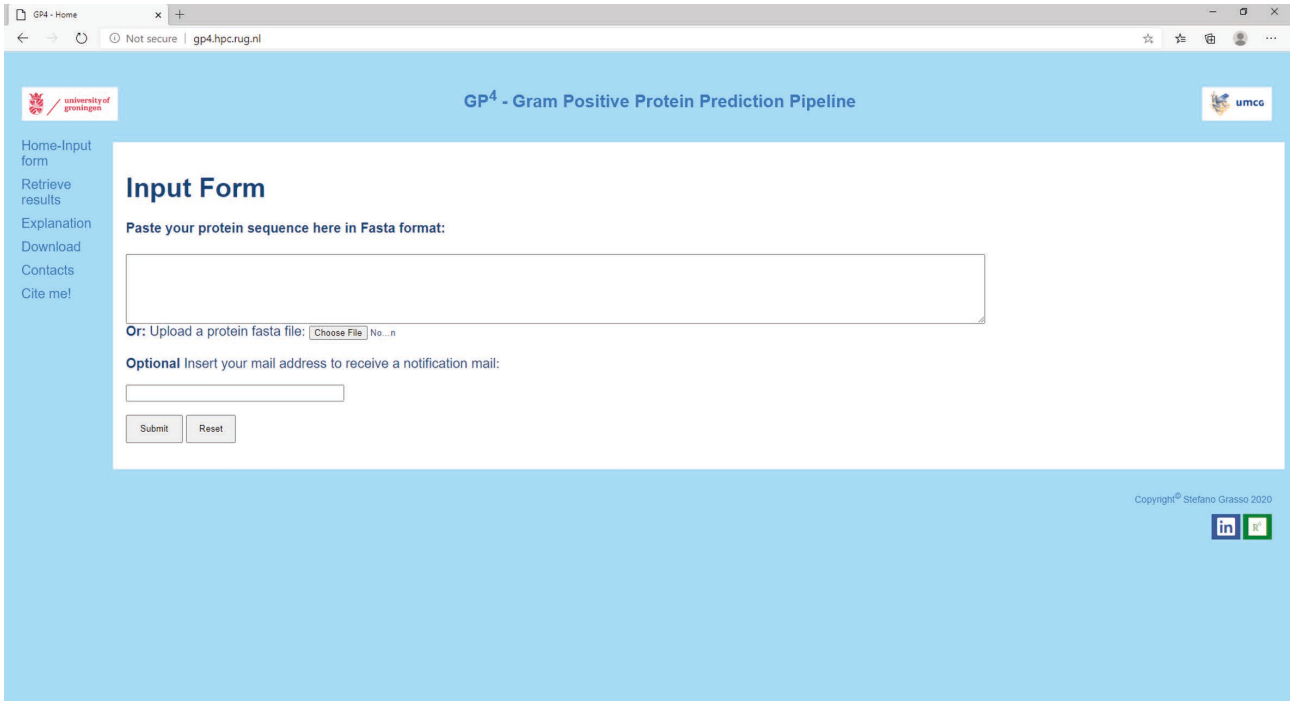


Figure 1. Homepage of the GP⁴ webserver. Input can either be pasted into the text box or uploaded as a file, both in fasta format. Optionally, it is possible to provide an email address to which the link with results will be sent. The interface is kept simple and no settings options are necessary. Results are stored for 7 days and can be retrieved at any moment through the specific page.

associated with the cytoplasmic side of the membrane, which pose a semantic challenge in regard to their localization [7], are predicted to be cytosolic. Further, we felt that membrane-bound proteins, like lipoproteins, should be classified on their own for the sake of clarity. To our knowledge, no other SCL prediction tool takes into account this issue, despite being discussed in literature [7,31,33]. Similarly, cell-wall proteins may be covalently attached to the peptidoglycan, or only transiently interact with it. In the absence of specific tools, we have tried to discriminate among these two possibilities. Furthermore, for extracellular proteins GP⁴ provides the most likely secretion pathway based on detected signals, taking into account not only the main Sec and Tat pathways, but also alternative ones such as: SecA2, the Wss route (i.e. the WXG100 secretion system, also called T7SSb), the flagellar export apparatus (FEA), the fimbrilin-protein exporter (FPE), and some lantibiotics and bacteriocins. Such aspects may be of lower relevance when analyzing bulk genomes for statistical purposes, but they may play major roles when analyzing specific protein candidates or engineered proteins. Lastly, it should be mentioned that GP⁴ fulfills the theorized properties of an expert system predictor [7,31], based on its high interpretability, explanatory power, and its accountability for synthetically designed proteins.

Software used

To develop the GP⁴ prediction pipeline, multiple candidate tools were evaluated to cover all relevant aspects, including: (i) detection of all possible secretion pathways; (ii) determination of TM topologies; and (iii) detection of domains, motifs, and repeats. For each aspect, the selection was further based on the reliability and efficiency of the various tools. Finally, usability and accessibility played a major role during selection. Considered criteria were the availability of downloadable or standalone versions, and limitations in the numbers or lengths of sequences that can be analyzed. Additionally, an overall parsimony approach was applied.

Signal peptides and secretion pathways

Detection and prediction of the correct secretion pathway is possibly one of the most challenging aspects of SCL prediction. The classical secretion pathway (Sec/signal peptidase I [SpI]) is the most studied and best characterized one and, thus, prediction of the respective SPs is most reliable. To detect these SPs, SignalP v. 4.1 [34], SignalP v. 5.0 [35], Phobius [36] and Predisi were exploited as they are specific for Sec SP detection. Additionally, also LipoP [37,38], despite being mainly designed for lipoprotein SPs (Sec/signal peptidase II [SpII]) can help to determine the type of secretion pathway. Similar to LipoP, also SignalP 5 has the ability to detect lipoprotein SPs, as well as Tat SPs (Tat/SpI). To complement this ability of SignalP 5, also TatP [39] was integrated in the GP^4 pipeline.

Unfortunately, tools to specifically predict other secretion pathways are currently not available. In particular, neither the signal peptides nor the proteins associated with protein secretion through ABC transporters, the SecA2 machinery, the FEA, FPEs, holins, the Wss route, and any other non-classical secretion system (including moonlighting proteins) can thus far be predicted with dedicated tools. To at least partially overcome this limitation, InterPro signatures peculiar to these classes of proteins have been exploited in GP^4 (see below).

Trans-membrane topology

TM helix detection is possibly one of the oldest predictable protein features. To detect them TMHMM [40] has been used in GP^4. Although TMHMM is a relatively old prediction tool, it is still considered efficient and reliable in its simplicity. To complement its ability to detect TM helices, topology predictions by Phobius were also taken into account.

Domains, motifs and repeats

For any other type of signal detection, signatures from InterPro [41] were used. InterPro collects and merges the entries from multiple databases and, additionally, manually curates them. This makes the various entries highly reliable. Nevertheless, given the fact that different methods and databases are used by InterProScan [42], different implemented signatures may have different levels of sensibility and sensitivity.

Manually curated lists of InterPro identifiers were created (Supplementary Table 1) for some of the main SCL targets, namely the secretion-associated signatures, Tat-associated signatures, lipoprotein-associated signatures, and cell-wall associated signatures (in turn, subdivided into covalent bonds, non-covalent bonds, and spore). In addition to those, and given the peculiar nature of some proteins, three additional lists were created to give a second, more detailed, level of SCL predictions: surface-associated signatures, pseudo-pilin- and fimbrilin-associated signatures, and short secreted peptide-associated signatures (e.g. lantibiotics, bacteriocins or similar). Of note, even though most of the selected signatures are known and widely used to associate proteins with SCLs, they are not officially associated to any SCL (more precisely any GO compartment).

To exploit the full InterPro potential, lists of specific GO terms were created (Supplementary Table 1). During the motif and domain analysis through the GO compartment field, InterProScan may detect some that are officially associated to a specific SCL.

Other included tools

Lastly, ProtCompB [43], an online predictor for bacterial SCLs, was added to the GP^4 pipeline for additional support in the decision-making process. ProtCompB combines several prediction methods, namely: ‘neural networks-based prediction, direct comparison with bases of homologous proteins of known localization,

comparisons of pentamer distributions calculated for query and database sequences, and prediction of certain functional peptide sequences, such as signal peptides and transmembrane segments' [43]. Thus, ProtCompB is fully complementary to the other afore-mentioned tools. In cases of doubtful decision-making, due to its highly reliable predictions [44,45], ProtCompB can help in steering the results in the right direction.

Discarded tools

Despite the availability of additional tools for certain specific tasks (e.g. TM/SP discrimination, or cell wall-binding predictions), it was decided to discard them, because these tools could not analyze more than one sequence at a time, the tool size would not be compatible with most users' machines, the outputs were graphical and would not be correctly parsed, or there were other usability issues.

Dataset

The GP4 prediction algorithm was designed based on the current state-of-the-art knowledge about protein sorting in Gram-positive bacteria, in particular in *Bacillus subtilis*. Accordingly, in detail testing was done based on the proteome of *B. subtilis* strain 168 (UP000001570).

Benchmark evaluation was performed with a test set, designated T1, of 374 proteins (summarized in Table 1; details of the dataset are presented in Supplementary Table 2). The dataset was built by retrieving from SwissProt [46], release 2020_02, all proteins belonging to the phyla of Firmicutes (id:1239) and Actinobacteria (id:1760) for which experimental evidence regarding the SCL was available for the respective species, and removing all proteins from in *B. subtilis* strain 168 (id:224308). This resulted in a set of 568 proteins. Afterwards, the redundancy was decreased to 25% identity by means of CD-HIT [47] (with standard settings), resulting in a total of 406 proteins. Finally, the dataset was further manually curated removing too short peptides or proteins whose SCL was classified as experimentally determined, but for which the published evidence was either poor or debatable. This yielded the final dataset of 374 proteins with known localization and at least one associated publication. Nevertheless, one should bear in mind that this curated dataset may still include some wrongly annotated proteins as there may be mistakes in the literature. This is exemplified by a public benchmarking dataset [21,48] where, amongst proteins classified as cytosolic, there are also some secreted proteins, such as Q933K8 (now P9WJD9 [49,50]) or P34020 [51]. Most likely, at the time of publication the respective secretion system was yet to be discovered. This underscores the need to perform benchmarking always on the latest state-of-art datasets.

Localization	Number of proteins
Cell Wall (CW)	58
Cytosolic (CYTO)	88
Extracellular (EXTRA)	133
Lipoproteins (LIPO)	9
Transmembrane (TM)	84
Multi-location (CW-TM)	2
Total	374

Table 1. Composition of the T1 protein dataset used for GP4 benchmarking.

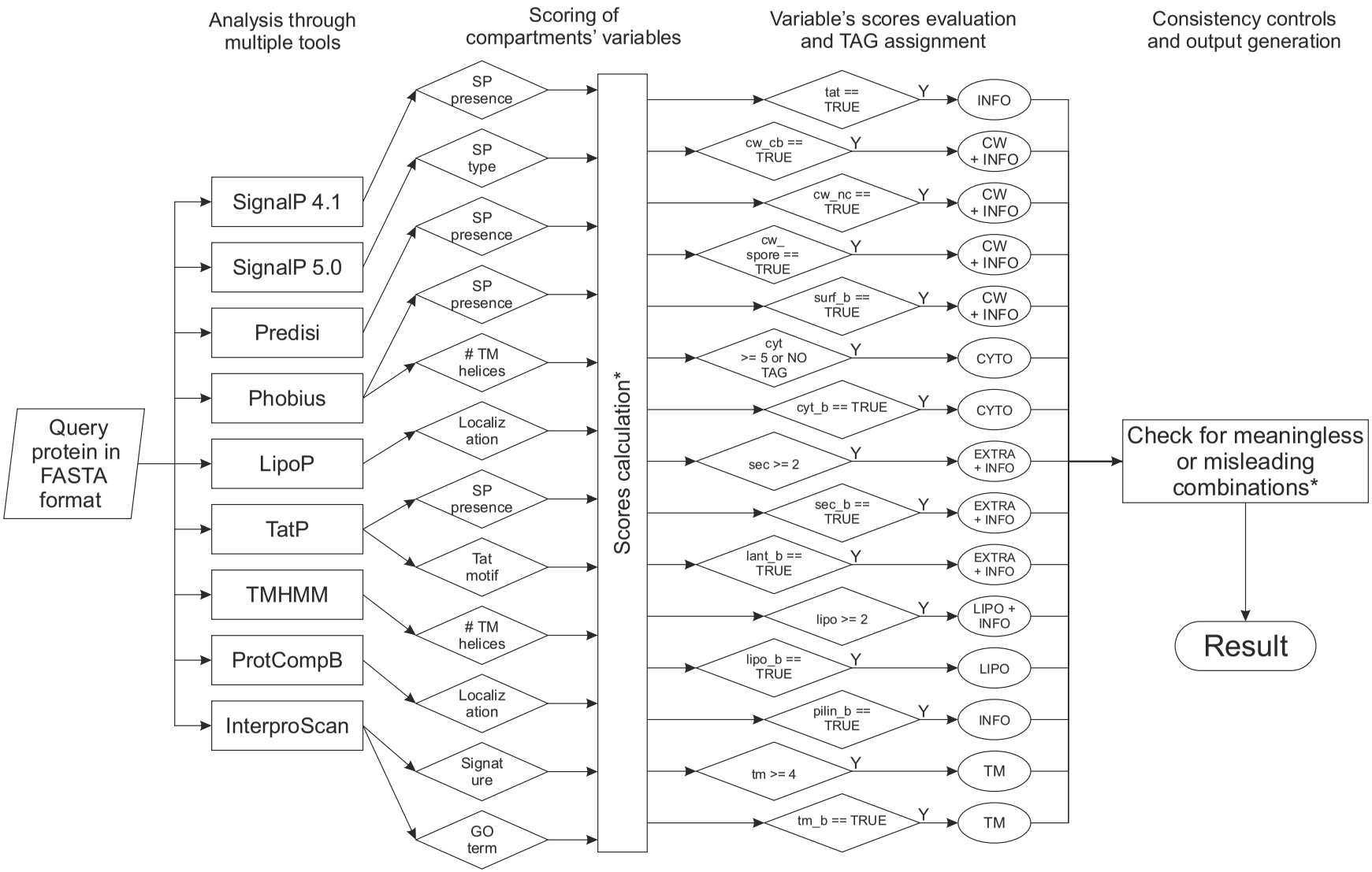


Figure 2. Summary of the prediction algorithm. The query protein is introduced in FASTA format and evaluated with multiple prediction tools. Based on the respective outputs, the values of 14 numerical and boolean variables are calculated. If certain conditions or thresholds are met, then specific TAGs plus additional information will be assigned to the query protein. Lastly, a check of the different assigned TAGs is performed in order to remove redundant information. If certain TAGs are in conflict, 'Unknown' is returned as a result. *, For the sake of clarity, particular details have been documented in Supplementary Table 3.

Implementation

The prediction algorithm, summarized in Figure 2 (for more details see also Supplementary Table 3), was written in Python v. 3.6. It combines the outputs of the above-mentioned tools with a simple scoring system to return a putative SCL. Results from the different exploited tools are parsed and, depending on each tool's output, scores for protein designation to the various compartments are increased or decreased. For each compartment's score there is a minimum threshold, which indicates the minimum amount of 'evidences' needed to assign an SCL TAG to a particular query protein. Additionally, if a sequence contains a compartment-related feature that indicates unequivocally a SCL, boolean variables can override the scores. This has been implemented particularly in regard to InterPro signatures. Finally, through a series of simple "if-then-else" conditionals, scores and boolean variables are combined in order to assign one, or more SCL TAGs to a protein. Whenever a compartment's score is higher than the set threshold, the respective SCL TAG is appended to the results for the query protein. The same has been implemented for boolean variables. Normally, multiple TAGs can be assigned, if each of them individually meets its own requirements. Nevertheless, a last consistency check is performed: TAG combinations are evaluated and, in some cases, modified either because they are meaningless or potentially misleading. In such cases, either redundant information is removed (e.g. 'EXTRA CW' becomes 'CW', as cell-wall proteins are intrinsically extracellular) or, when in conflict, 'Unknown' is returned as a result (e.g. 'EXTRA CYTO' becomes 'Unknown', as there is no combination of signals that could lead to such a prediction). Additionally, 'Unknown' is also returned if no score reaches the required threshold. Next to the main SCL prediction, GP⁴ provides additional information, such as the secretion pathway used by a particular secretory protein, the signal peptidase cleavage site of putative SPs, or the anticipated type of interaction with the cell wall. Such information is provided either paired with a specific SCL (e.g. detection of an LPXTG motif for covalent protein attachment to the cell wall), or independently (e.g. a Tat motif or pilin-like motif, as these motifs may suggest a final SCL, but do not completely determine it).

Evaluation method

The test set T1 was used to evaluate the current prediction method and to compare it with other tools, namely: PsortB [19] v. 3, LocTree3 [52], pLoc_bal-mGpos [21], Cello v. 2.5 [20], and BUSCA [53]. Proteins were analyzed and predictions were used to calculate sensitivity, specificity, precision, accuracy and the Matthews correlation coefficient (MCC) for each class of proteins. These parameters are defined as follows:

- Sensitivity: $\frac{TP}{TP+FN}$
- Specificity: $\frac{TN}{TN+FP}$
- Precision: $\frac{TP}{TP+FP}$
- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- MCC: $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+PF)(TN+FN)}}$

TP, FP, TN and FN indicate true positives, false positives, true negatives and false negatives, respectively, for each localization. For a perfect set of predictions, the MCC value is 1, for a completely random prediction it will be 0 and for a perfect reverse prediction the MCC is -1. Lastly, the overall values for each tool were calculated as the weighted average between the various classes.

Results & Discussion

Benchmark of SCL predictions

Benchmark comparative analyses with GP⁴ were performed using the majority of currently existing tools,

normally returning alternate results [5,53–59]. This was to be expected considering the fact that the composition of the train and test sets [60] and the relative internal amino acid sequence similarity levels [15,20] will have a major impact on the outcome. Most tools tend to be more precise the more the query sequence, or the species from which it is derived, are related to the elements incorporated in the training set. Even more so, this bias is present in homology-based prediction tools that rely on the presence and correct annotation of proteins within the respective database. Nevertheless, also tools based on the identification of motifs and signatures can have similar biases. This possibility should therefore be considered with respect to the following sections of this paper on the performance of GP^4.

In addition to GP^4, five other SCL predictors, namely PsortB v. 3 [19], LocTree3 [52], pLoc_bal-mGpos [21], CELLO v. 2.5 [20], and BUSCA [53] were benchmarked on the T1 dataset of 374 protein with known localization. The overall results are summarized in Table 2. LocTree3 turned out to be the best-performing tool with an overall MCC of 0.760. Nevertheless, it should be noted that LocTree3 is not able to discriminate between proteins from Gram-positive and Gram-negative bacteria. Additionally, it is not able to predict the cell wall as a compartment, classifying cell wall proteins simply as extracellular. Similarly, also BUSCA can only predict three compartments in Gram-positive bacteria, lacking the cell wall class. Similar to GP^4, BUSCA is based on combining multiple tools for the detection of sorting signals but, unfortunately, for Gram-positive bacteria only SPs and TM helices are searched. Lacking many of the known bacterial sorting signals, BUSCA performs worse than LocTree3 with an MCC of 0.625, but it still provides useful information about the position of potential TM helices in detected SPs. Due to their simplicity, both tools can be an interesting choice to obtain a broad idea of the overall distribution of proteins. Nevertheless, in case of querying single proteins, or when a high level of precision is needed, more suitable tools are available. In particular, the here presented GP^4 prediction tool, together with PsortB, pLoc_bal-mGpos, and CELLO are more suitable for a comparison as they include the four main SCLs of Gram-positive bacteria. Of note, GP^4 provides an extra prediction result, namely 'LIPO' for lipoproteins, which does not in itself represent a sub-cellular compartment, but predicts with striking precision the membrane association of such proteins, resulting in an MCC of 1. Among these four tools, pLoc_bal-mGpos performed strikingly worse than expected with an overall MCC of 0.349. In contrast, CELLO proved to be better overall, but predicted only one cell-wall protein in the whole dataset, which lowered the overall scores. Given these results, it would make more sense to use simpler tools, like BUSCA or LocTree3, which can deliver better overall predictions. GP^4, instead, performed slightly better than PsortB, with respective MCCs of 0.709 and 0.698. This outcome for PsortB was comparable with previous benchmarking analyses [54,61]. It must be noted that, despite the similar MCC values, PsortB predicted 17.11% of the proteins as unknown, while for GP^4 only 3.74% were predicted as unknown. More in detail, GP^4 turned out the best-performing tool among the tested ones, with an MCC of 0.670, for cell-wall proteins (0.574 for PsortB; see also Supplementary Table 4). Instead, PsortB performed apparently better for extracellular proteins with an MCC of 0.736 (0.615

Tool	Sensitivity	Specificity	Precision	Accuracy	MCC
GP^4	0.78	0.91	0.82	0.88	0.71
PsortB	0.77	0.92	0.81	0.78	0.70
LocTree3	0.83	0.93	0.86	0.92	0.76
pLoc_bal-mGpos	0.52	0.83	0.53	0.77	0.35
BUSCA	0.71	0.91	0.82	0.79	0.63
CELLO v. 2.5	0.70	0.86	0.75	0.83	0.55

Table 2. Summary of the GP4 benchmark analysis. The Table summarizes the sensitivity, specificity, precision, accuracy and MCC for all benchmarked tools.

for GP⁴), but it must be noted that these values are influenced by the relevant difference in the rates of 'unknown' predictions for secreted proteins, namely 25.56% for PsortB versus 0.75% for GP⁴. This makes GP⁴ the best option to predict extracellular proteins in absolute numbers (i.e. taking into account the 'unknown' predictions by the two tools), as well as the most accurate. The improvement gained by GP⁴ for the prediction of extracellular and cell-wall proteins can probably be attributed to the detection of specific compartment-related domains.

Usage on modified proteins

If SCL prediction tools would be classified as text-editors, PsortB would be considered as a WYSIWYM (what you see is what you mean), because it returns the SCL for the specific class of proteins. On the contrary GP⁴ would be considered as a WYSIWYG (what you see is what you get) tool that predicts only what can be directly evinced from the actual query sequence. This is best exemplified by barnases, which are extracellular ribonucleases produced by various *Bacillus* species. As most secreted proteins, barnases do possess a Sec SP necessary for their export. The reference barnase was first discovered in *Bacillus amyloliquefaciens* (P00648) and possesses a SP and a propeptide according to SwissProt. Among homologous proteins with at least 90% identity, there is another barnase from *Bacillus circulans* (P35078). According to the annotation, this protein is 47 residues shorter and lacks a SP (Figure 3). The apparent lack of a SP is probably to be attributed to misannotation or low-quality sequence assembly, although we were not able to retrieve a SP from the corresponding nucleotide sequence (data not shown). However, the protein in this form, i.e. without a SP, is unlikely to be secreted by a Gram-positive bacterium. Similarly, if we were to produce such a truncated protein in a heterologous strain, e.g. *B. subtilis*, it would hardly be secreted. Yet, all the tested prediction tools designated both barnases of *B. amyloliquefaciens* and *B. circulans* as extracellular proteins. This is formally correct for the regular barnases, but not for the barnase of *B. circulans* as it was annotated. On the contrary, GP⁴ predicted the *B. amyloliquefaciens* barnase to be secreted via Sec, while the truncated barnase of *B. circulans* was predicted to be cytoplasmic, as the amino acidic sequence lacks a SP. The latter approach is certainly favorable in the context of engineered organisms, but may be misleading when annotating wild-type genomes, and it can certainly not compensate for annotation errors. The latter can instead be managed by other approaches.

The main consequence of a WYSYWIG approach is the impossibility to predict protein sorting based on unknown or poorly characterized pathways. This should not be regarded as a negative aspect, but rather an incentive to improve the current knowledge and understanding of bacterial sorting mechanisms and, at the same time, to develop novel and more precise tools to detect specific sorting signals embedded in the

```
P00648     1 MMKMEGIALKKRLSWISVCLLVLVSAAGMLFSTAAKTETSSHKAHTEAQVINTFDGVADY
P35078     1 -----------------------------------------------AQVINTFDGVADY

P00648    61 LQTYHKLPDNYITKSEAQALGWVASKGNLADVAPGKSIGGDIFSNREGKLPGKSGRTWRE
P35078    14 LLTYHKLPDNYITKSEAQALGWVASKGNLADVAPGKSIGGDIFSNREGKLPAKSGRTWRE

P00648   121 ADINYTSGFRNSDRILYSSDWLIYKTTDHYQTFTKIR
P35078    74 ADINYTSGFRNSDRILYSSDWLIYKTTDHYKTFTKIR
```

Figure 3. Sequence alignment of barnases from B. amiloliquefaciens and B. circulans. The barnase sequences were aligned with Clustal Omega. Gray shading marks the SP of the B. amyloliquefaciens barnase (P00648), which is absent from the barnase of B. circulans (P35078). Depending on the selected SCL prediction approach, P35078 would either be assigned as an extracellular protein since it belongs to a family of extracellular RNases, or as a cytosolic protein since it lacks a SP. Clearly, in absence of an appropriate SP, bacterial export of a protein with the P35078 sequence is unlikely.

amino acidic sequence. In fact, in the present study we show how SCL prediction based on detectable sorting signals can be more powerful than other approaches, regardless of the fact that many signals or motifs are still to be discovered or elucidated with respect to their function.

Conclusions & Future Perspectives

In conclusion, the here presented GP^4 tool seems to perform better than other SCL predictors, despite its intrinsic inability to predict SCLs for proteins that follow poorly characterized sorting pathways. In particular, GP^4 should be appropriate for synthetic organisms, or organisms with little studied genomes. Furthermore, we consider GP^4 the most widely-applicable tool for SCL predictions in Gram-positive bacteria. Due to its superiority in detecting extracellular and cell-wall proteins, it can probably help in the identification of novel targets for drugs against pathogenic Firmicutes and Actinobacteria. This is a consequence of its design, where prior knowledge on genomes or proteins is not necessary. On the other hand, the applicability of GP^4 is limited by our overall understanding of protein sorting. For instance, GP^4 was proven effective for SCL predictions in Actinobacteria, but it cannot predict the outer membrane proteins of this group. Only PsortB 3.0 can predicted such outer membrane proteins, but only through a homology-based approach, as there is currently no other method or tool to detect this class of proteins. GP^4 will thus predict Actinobacterial outer membrane proteins as secreted proteins, and it will remain a task for the user to perform further analyses to correctly assign their SCL. Altogether, we anticipate that experienced users will find GP^4 applicable also for SCL predictions in other less-studied organisms, such as Tenericutes, but due to the current lack of proteins with known localization we have not tested this.

Particular attention should be attributed to the development of SCL prediction tools. While various tools have thus far been developed, none of them proved to be truly superior. Therefore, we advocate a paradigm shift in the development of SCL predictors. It was already known that meta-predictors perform better than single-purpose predictors [54,62–64], because the meta-predictors exploit specific strengths while compensating for weaknesses of the individual tools. Yet, few advancements have been made in this direction, and no meta-predictor webserver for Gram-positive was thus far available. At least in prokaryotes, a stronger effort in developing sorting signal detectors, analogous to SignalP, should be made. In this regard also the usability and parsability should be taken into account. This will lead to the creation of tools with standalone versions that do not rely exclusively on centralized webservers, and with standardized outputs that are easy to programmatically read and parse. These are prerequisites to develop better and more efficient meta-predictors, which could even be presented in a modular form with different tools being loaded, depending on the scope or source of the query. With these premises the future development of SCL predictors may be brought to superior levels, as was achieved for the other two classes of functional annotation [65–69], and in other fields [70,71].

Key points

- Multiple methods for protein subcellular localization prediction are available, with different advantages and disadvantages depending on the origin of the query sequence.
- We propose to combine multiple single-feature predictors to mimic protein sorting within Gram-positive bacterial cells. This approach is knowledge-based and relies on our current understanding of prokaryotic biology, but not on prior knowledge of closely related organisms.
- GP⁴ is the first tool which encompasses the capability to predict: 1) non-canonically secreted proteins; 2) lipoproteins, 3) cell-wall binding and interacting domains.
- When benchmarked against other subcellular localization prediction tools, the presented GP⁴ outperforms the other tools. In addition, GP⁴ provides extra information regarding the subcellular localization of query proteins, provides all data used to draw such conclusion, and allows for a re-interpretation of results by experienced users.
- A webserver running GP⁴ is available at: http://gp4.hpc.rug.nl/

Code availability

In addition to the webserver, a standalone version of GP⁴ is available at: https://github.com/grassoste/GP4_standalone

Funding

This work was supported by the European Union's Horizon 2020 Program, Marie Skłodowska-Curie Actions (MSCA), under REA grant agreement no. 642836.

Acknowledgments

We would like to thank the developers of all prediction tools mentioned in this paper. Without the software they developed, the presented GP⁴ tool could not exist. We also thank the team of CIT, in particular Cristian Marocico, Egon Rijpkema, and Fokke Dijkstra, who supported us in server implementation.

References

1. Nielsen H, Tsirigos KD, Brunak S, et al. A Brief History of Protein Sorting Prediction. Protein J. 2019; 38:200–216

2. Nielsen H. Protein sorting prediction. Methods Mol. Biol. 2017; 1615:23–57

3. Nielsen H. Predicting subcellular localization of proteins by bioinformatic algorithms. Curr. Top. Microbiol. Immunol. 2017; 404:129–158

4. Wan S, Mak M-W. Machine Learning for Protein Subcellular Localization Prediction. Berlin: de Gruyter, 2015

5. Gardy JL, Brinkman FSL. Methods for predicting bacterial protein subcellular localization. Nat. Rev. Microbiol. 2006; 4:741–751

6. Dönnes P, Höglund A. Predicting Protein Subcellular Localization: Past, Present, and Future. Genomics. Proteomics Bioinformatics 2004; 2:209–215

7. Nakai K, Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria. Proteins Struct. Funct. Genet. 1991; 11:95–110

8. Jones CE, Brown AL, Baumann U. Estimating the annotation error rate of curated GO database sequence annotations. BMC Bioinformatics 2007; 8:170

9. Perdigão N, Heinrich J, Stolte C, et al. Unexpected features of the dark proteome. Proc. Natl. Acad. Sci. U. S. A. 2015; 112:15898–903

10. Valencia A. Automatic annotation of protein function. Curr. Opin. Struct. Biol. 2005; 15:267–274

11. Kumar D, Kumar Mondal A, Kutum R, et al. Proteogenomics of rare taxonomic phyla: A prospective treasure trove of protein coding genes. Proteomics 2016; 16:226–240

12. Lobb B, Tremblay BJM, Moreno-Hagelsieb G, et al. An assessment of genome annotation coverage across the bacterial tree of life. Microb. Genomics 2020; 6:e000341

13. Gilks WR, Audit B, De Angelis D, et al. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics 2002; 18:1641–1649

14. Gilks WR, Audit B, De Angelis D, et al. Percolation of annotation errors through hierarchically structured protein sequence databases. Math. Biosci. 2005; 193:223–234

15. Nair R, Rost B. Sequence conserved for subcellular localization. Protein Sci. 2009; 11:2836–2847

16. Imai K, Nakai K. Prediction of subcellular locations of proteins: Where to proceed? Proteomics 2010; 10:3970–3983

17. Devos D, Valencia A. Practical limits of function prediction. Proteins Struct. Funct. Bioinforma. 2000; 41:98–107

18. Addou S, Rentzsch R, Lee D, et al. Domain-Based and Family-Specific Sequence Identity Thresholds Increase the Levels of Reliable Protein Function Transfer. J. Mol. Biol. 2009; 387:416–430

19. Yu NY, Wagner JR, Laird MR, et al. Sequence analysis PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. 2010; 26:1608–1615

20. Yu C-S, Chen Y-C, Lu C-H, et al. Prediction of protein subcellular localization. Proteins Struct. Funct. Bioinforma. 2006; 64:643–651

21. Xiao X, Cheng X, Chen G, et al. pLoc_bal-mGpos: Predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. Genomics 2019; 111:886–892

22. Lu Z, Szafron D, Greiner R, et al. Predicting subcellular localization of proteins using machine-learned classifiers. Bioinformatics 2004; 20:547–556

23. Danchin A, Ouzounis C, Tokuyasu T, et al. No wisdom in the crowd: genome annotation in the era of big data - current status and future prospects. Microb. Biotechnol. 2018; 11:588–605

24. Promponas VJ, Iliopoulos I, Ouzounis CA. Annotation inconsistencies beyond sequence similarity-based

function prediction – phylogeny and genome structure. Stand. Genomic Sci. 2015; 10:108
25. Kyrpides NC, Ouzounis CA. Errors in genome reviews. Science 1998; 281:1457
26. Pallen M, Wren B, Parkhill J. 'Going wrong with confidence': misleading sequence analyses of CiaB and ClpX. Mol. Microbiol. 1999; 34:195–195
27. Krishnappa L, Dreisbach A, Otto A, et al. Extracytoplasmic proteases determining the cleavage and release of secreted proteins, lipoproteins, and membrane proteins in *Bacillus subtilis*. J. Proteome Res. 2013; 12:4101–4110
28. Desvaux M, Hébraud M, Talon R, et al. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. Trends Microbiol. 2009; 17:139–145
29. Megrian D, Taib N, Witwinowski J, et al. One or two membranes? Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. Mol. Microbiol. 2020; 113:659–671
30. Zuber BB, Haenni M, Ribeiro T, et al. Granular Layer in the Periplasmic Space of Gram-Positive Bacteria and Fine Structures of Enterococcus gallinarum and Streptococcus gordonii Septa Revealed by Cryo-Electron Microscopy of Vitreous Sections. J. Bacteriol. 2006; 188:6652–6660
31. Horton P, Mukai Y, Nakai K. Protein subcellular localization prediction. Pract. Bioinformatician 2004; 193–216
32. Danchin A, Fang G. Unknown unknowns: essential genes in quest for function. Microb. Biotechnol. 2016; 9:530–540
33. Nakai K, Kanehisa M. A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 1992; 14:897–911
34. Petersen TN, Brunak S, Von Heijne G, et al. SignalP 4.0: Discriminating signal peptides from transmembrane regions. Nat. Methods 2011; 8:785–6
35. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. 2019; 37:420–423
36. Käll L, Krogh A, Sonnhammer ELL. A Combined Transmembrane Topology and Signal Peptide Prediction Method. J. Mol. Biol. 2004; 338:1027–1036
37. Rahman O, Cummings SP, Harrington DJ, et al. Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of Gram-positive bacteria. World J. Microbiol. Biotechnol. 2008; 24:2377–2382
38. Juncker AS, Willenbrock H, von Heijne G, et al. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. 2003; 12:1652–1662
39. Bendtsen JD, Nielsen H, Widdick D, et al. Prediction of twin-arginine signal peptides. BMC Bioinformatics 2005; 6:167
40. Krogh a, Larsson B, von Heijne G, et al. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 2001; 305:567–580
41. Mitchell AL, Attwood TK, Babbitt PC, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019; 47:D351–D360
42. Jones P, Binns D, Chang H-Y, et al. InterProScan 5: genome-scale protein function classification. 2014; 30:1236–1240
43. ProtCompB - Predict the sub-cellular localization of bacterial proteins. http://www.softberry.com/berry.phtml?topic=pcompb&group=programs&subgroup=proloc
44. Taheri-Anganeh M, Khatami SH, Jamali Z, et al. In silico analysis of suitable signal peptides for secretion of a recombinant alcohol dehydrogenase with a key role in atorvastatin enzymatic synthesis. Mol. Biol. Res. Commun. 2019; 8:17–26
45. Mohammadi S, Mostafavi-Pour Z, Ghasemi Y, et al. In silico Analysis of Different Signal Peptides for the Excretory Production of Recombinant NS3-GP96 Fusion Protein in Escherichia coli. Int. J. Pept. Res. Ther.

2019; 25:1279–1290
46. UniProt Consortium. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019; 47:D506–D515
47. Huang Y, Niu B, Gao Y, et al. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010; 26:680–2
48. Xiao X, Cheng X, Su S, et al. pLoc-mGpos: Incorporate Key Gene Ontology Information into General PseAAC for Predicting Subcellular Localization of Gram-Positive Bacterial Proteins. Nat. Sci. 2017; 09:330–349
49. McLaughlin B, Chon JS, MacGurn JA, et al. A mycobacterium ESX-1-secreted virulence factor with unique requirements for export. PLoS Pathog. 2007; 3:1051–1061
50. Chen JM, Zhang M, Rybniker J, et al. Mycobacterium tuberculosisEspB binds phospholipids and mediates EsxA-independent virulence. Mol. Microbiol. 2013; 89:1154–1166
51. Croux C, Canard B, Goma G, et al. Autolysis of Clostridium acetobutylicum ATCC 824. J. Gen. Microbiol. 1992; 138:861–869
52. Goldberg T, Hecht M, Hamp T, et al. LocTree3 prediction of localization. Nucleic Acids Res. 2014; 42:W350-5
53. Savojardo C, Martelli PL, Fariselli P, et al. BUSCA: an integrative web server to predict subcellular localization of proteins. Nucleic Acids Res. 2018; 46:459–466
54. Magnus M, Pawlowski M, Bujnicki JM. MetaLocGramN: A meta-predictor of protein subcellular localization for Gram-negative bacteria. Biochim. Biophys. Acta - Proteins Proteomics 2012; 1824:1425–1433
55. Hochreiter S, Heusel M, Obermayer K. Fast model-based protein homology detection without alignment. Bioinformatics 2007; 23:1728–1736
56. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, et al. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017; 33:3387–3395
57. Orioli T, Vihinen M. Benchmarking subcellular localization and variant tolerance predictors on membrane proteins. BMC Genomics 2019; 20:547
58. Kho CW, Park SG, Cho S, et al. Confirmation of Vpr as a fibrinolytic enzyme present in extracellular proteins of *Bacillus subtilis*. Protein Expr. Purif. 2005; 39:1–7
59. Sperschneider J, Catanzariti A-M, DeBoer K, et al. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Sci. Rep. 2017; 7:44598
60. Zhou H, Yang Y, Shen H-B. Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. Bioinformatics 2017; 33:843–853
61. Paramasivam N, Linke D. Clubsub-P: Cluster-based subcellular localization prediction for gram-negative bacteria and archaea. Front. Microbiol. 2011; 2:218
62. Liu J, Kang S, Tang C, et al. Meta-prediction of protein subcellular localization with reduced voting. Nucleic Acids Res. 2007; 35:96
63. Lertampaiporn S, Nuannimnoi S, Vorapreeda T, et al. PSO-LocBact: A Consensus Method for Optimizing Multiple Classifier Results for Predicting the Subcellular Localization of Bacterial Proteins. Biomed Res. Int. 2019; 2019:5617153
64. Hooper CM, Tanz SK, Castleden IR, et al. Data and text mining SUBAcon: a consensus algorithm for unifying the subcellular localization data of the Arabidopsis proteome. 2014; 30:3356–3364
65. Zielezinski A, Dziubek M, Sliski J, et al. ORCAN - A web-based meta-server for real-time detection and functional annotation of orthologs. Bioinformatics 2017; 33:1224–1226
66. Griesemer M, Kimbrel JA, Zhou CE, et al. Combining multiple functional annotation tools increases coverage of metabolic annotation. BMC Genomics 2018; 19:948
67. Friedberg I, Harder T, Godzik A. JAFA: a protein function annotation meta-server. Nucleic Acids Res. 2006;

34:W379-81

68. Pereira C, Denise A, Lespinet O. A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics 2014; 15 Suppl 6:S16

69. Reijnders MJMF. CrowdGO: a wisdom of the crowd-based Gene Ontology annotation tool. bioRxiv 2019; 731596

70. Manavalan B, Basith S, Shin TH, et al. MAHTPred: A sequence-based meta-predictor for improving the prediction of anti-hypertensive peptides using effective feature representation. Bioinformatics 2019; 35:2757–2765

71. Kara A, Vickers M, Swain M, et al. Genome-wide prediction of prokaryotic two-component system networks using a sequence-based meta-predictor. BMC Bioinformatics 2015; 16:1–9

CHAPTER 3

GINGIMAPS: PROTEIN LOCALIZATION IN THE ORAL PATHOGEN *PORPHYROMONAS GINGIVALIS*

Giorgio Gabarrini, Stefano Grasso,
Arie Jan van Winkelhoff, Jan Maarten van Dijl

Microbiology and Molecular Biology Reviews, 2020, 84(1):e00032-19

Summary

Porphyromonas gingivalis is an oral pathogen involved in the widespread disease periodontitis. In recent years, however, this bacterium has been implicated in the etiology of another common disorder, the autoimmune disease rheumatoid arthritis. Periodontitis and rheumatoid arthritis were known to correlate for decades, but only recently a possible molecular connection underlying this association has been unveiled. *P. gingivalis* possesses an enzyme that citrullinates certain host proteins and, potentially, elicits autoimmune antibodies against such citrullinated proteins. These autoantibodies are highly specific for rheumatoid arthritis and have been purported both as symptom and potential cause of the disease. The citrullinating enzyme and other major virulence factors of *P. gingivalis*, including some that were implicated in the etiology of rheumatoid arthritis, are targeted to the host tissue as secreted or outer membrane-bound proteins. These targeting events play pivotal roles in the interactions between the pathogen and its human host. Accordingly, the overall protein sorting and secretion events in *P. gingivalis* are of prime relevance for understanding its full disease-causing potential and for developing preventive and therapeutic approaches. The aim of this review is therefore to offer a comprehensive overview of the subcellular and extracellular localization of all proteins in three reference strains and four clinical isolates of *P. gingivalis*, as well as the mechanisms employed to reach these destinations.

Supplementary Files

The Supplementary Files can be accessed at: https://bitbucket.org/teto1991/gingimaps.

Supplementary files avaialble at: https://github.com/grassoste/Thesis-supplementary-files

Introduction

Porphyromonas gingivalis is a Gram-negative, black-pigmented anaerobic bacterium belonging to the Bacteroidetes phylum. Although this bacterium is often described as rod-shaped, its appearance is more reminiscent of a little sausage (Fig. 1). *P. gingivalis* initially garnered interest as a model organism for bacteria of the *Cytophaga-Flavobacterium-Bacteroides* (CFB) group and, later on, as an oral pathogen. The focus on *P. gingivalis* has recently peaked with the discovery of a new protein secretion system (1) and with the evidence of its involvement in Alzheimer's disease (2). However, this bacterium is best known as a major etiological agent of the oral disease periodontitis (3, 4), being present in almost 85% of severe cases (5-7). Periodontitis is an inflammatory disorder affecting the tissue surrounding the teeth, the periodontium, potentially leading to tooth loss. Severe forms of periodontitis have a global prevalence of ~11%. However, depending on the degree of severity, socio-economic status and oral hygiene, this disease can affect up to 57% of particular populations (8, 9). In the USA, for example, 46% of adults are affected by this disorder, with 8.9% presenting severe forms (9). This extremely high incidence establishes periodontitis as one of the most common diseases, and as the main cause of tooth loss worldwide (9, 10).

Interestingly, periodontitis has been associated with several health conditions, such as diabetes, heart diseases, Alzheimer's disease, and rheumatoid arthritis (RA). In the case of diabetes, a two-way relationship was proposed, where the inflammatory mediators released in response to a periodontal infection would have an adverse effect on glycemic control, while diabetes-driven factors, such as impaired chemotaxis, reduced collagen synthesis, and increased collagenase production would, in turn, enhance the severity of periodontitis (11-16). The association between periodontitis and heart diseases, on the other hand, is more tenuous than the one with diabetes, and no potential mechanistic links are currently known (17-19). Investigations on the association of periodontitis with dementia support the potential involvement of periodontitis in this cognitive disorder both at the immunomodulatory level, which would relate to the systemic inflammatory responses caused by this oral disease, and at the physiological level, which could relate to possible micronutrient deficiencies (e.g. for thiamine and vitamin B12) that may arise from dietary changes as a consequence of tooth loss and that potentially lead to cognitive impairment (20, 21). A special case has been made for the most common type of dementia, Alzheimer's disease, where *P. gingivalis* has been proposed to play a significant role (2, 20-23). In particular, it was suggested that the secretion of particular cysteine proteases called gingipains may cause neuronal damage, which would be supported by the fact that these proteases, along with bacterial DNA, were detected in the brains of Alzheimer's disease patients (2). Lastly, the association of periodontitis with RA has been studied most intensively (24-42). RA is an inflammatory autoimmune disorder of which the etiology is still not fully understood, and that is clinically associated with periodontitis. In several countries, the prevalence of periodontitis was reported to be increased among RA patients in comparison with the general population (24, 29, 36, 37, 40, 43, 44). Correspondingly, RA was found to be more prevalent among patients with periodontitis (35-37, 40, 44), which supports the hypothesis that an intimate connection exists between the two disorders.

The suspected role of *P. gingivalis* in the interplay between periodontitis and RA has drawn attention to the bacterium's citrullinating enzyme (25, 27, 28, 32, 34, 41). This enzyme, a peptidylarginine deiminase (PAD), catalyzes the conversion of arginine into citrulline residues in a post-translational protein modification called citrullination. Citrullination has the potential to alter the net charge of a substrate protein, possibly leading to severe changes in its structure and function (27). Although citrullination is a physiological process that takes place in a wide variety of healthy tissues as a general regulatory mechanism, especially during apoptosis, it is also associated with inflammatory processes.

While peptidylarginine deiminases are highly conserved in mammals, only three bacteria of the genus *Porphyromonas* are known to produce such enzymes (27, 38, 45-47). The PAD of *P. gingivalis* (PPAD)

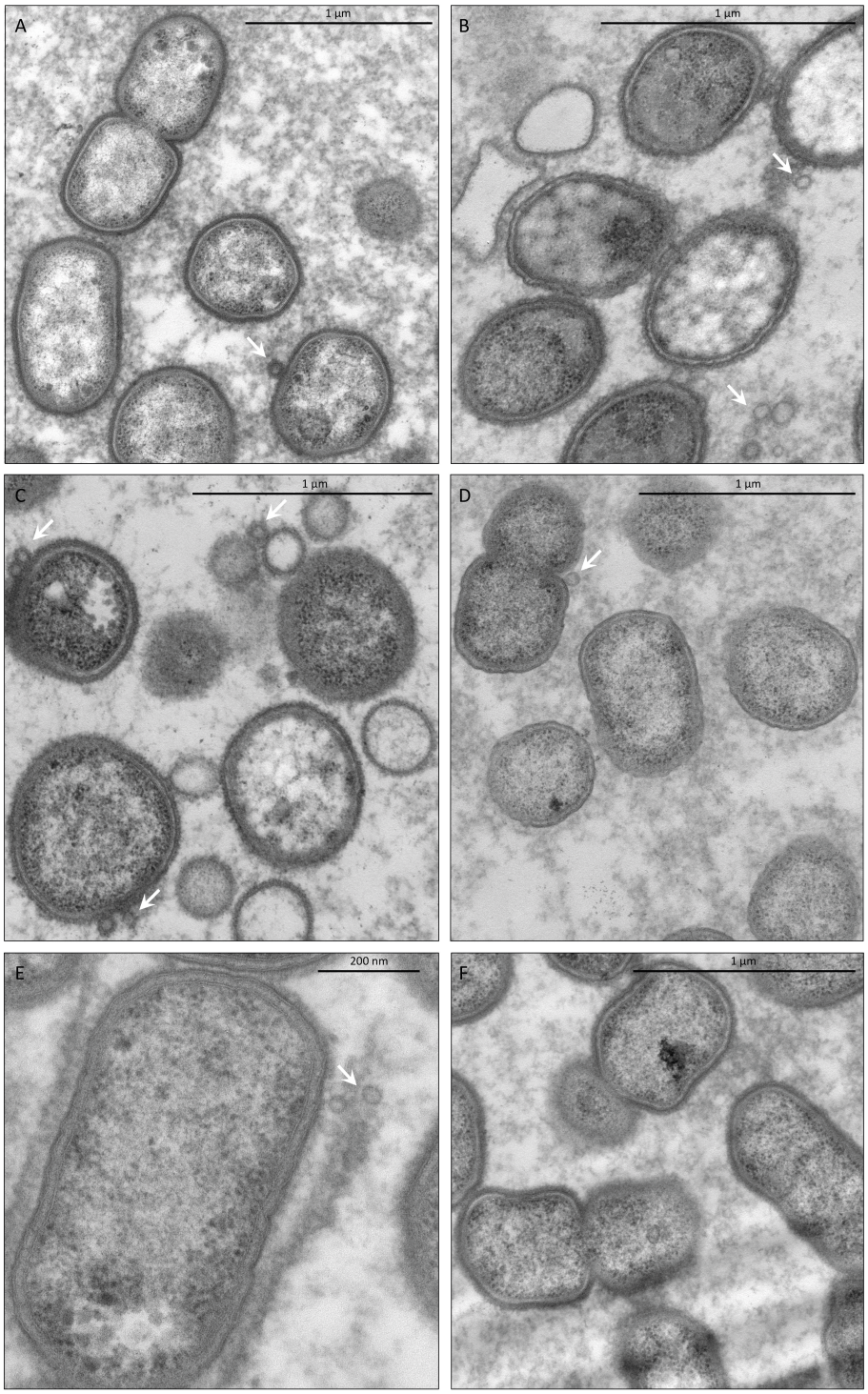


Figure 1. ***Porphyromonas gingivalis.*** Electron micrographs of *P. gingivalis* **A)** type strain W83, and the clinical strains **B)** 505700, **C)** 512915, **D)** 505759, and **E-F)** MDS33. Note the capture of OMV formation in panels A, B, C and E as marked by white arrows.

and the homologous enzymes from *Porphyromonas loveana* and *Porphyromonas gulae* share no evolutionary relationship with the mammalian PADs (47, 48). Remarkably, PPAD is believed to citrullinate certain human host proteins that, especially in genetically predisposed subjects (27, 34), can stimulate the production of anti-citrullinated protein antibodies (ACPAs) (27, 31, 38, 39, 45). These ACPAs have 95% specificity and 68% sensitivity for RA (49, 50). Interestingly, like many other bacterial virulence factors, the citrullinating enzyme of *P. gingivalis* is targeted to the host milieu. In particular, PPAD was detected in gingival tissue of patients with severe periodontitis (120). *In vitro* studies have shown that *P. gingivalis* secretes PPAD both in a secreted soluble state and in an outer membrane vesicle (OMV)-bound state (45, 51, 52). In addition, a substantial portion of PPAD remains associated with the bacterial cell in an outer membrane (OM)-bound state. Other proteins that play a role in *P. gingivalis* colonization of the periodontal pockets, such as hemagglutinins and fimbrial components, or that have been implicated in RA etiology, such as several cysteine proteases, are exposed on the bacterial cell surface (53). These findings and the possible roles of *P. gingivalis* in RA and other diseases focus interest on the mechanisms and pathways responsible for protein sorting and export in this bacterium. In this context it is noteworthy that *P. gingivalis* is an extremely successful oral pathogen that does not only take advantage of proteinaceous virulence factors, but also of non-proteinaceous virulence factors, such as capsule (54, 55) and lipopolysaccharides (56), to manifest itself as a “keystone” species within subgingival biofilms (57, 58).

All factors that contribute to the success of *P. gingivalis* in the periodontal pockets, where it mainly resides, are contained in the bacterium’s proteome. Importantly, the presence of particular bacterial proteins in a specific subcellular compartment is related to their biological function. Secreted proteins, for example, are involved in processes that take place at the cell surface or beyond, such as nutrient acquisition, cell motility, cell-cell communication or host colonization and invasion. Accordingly, bacterial cell surface proteins represent excellent targets for drugs or vaccines (160). Moreover, knowledge of the subcellular localization of proteins is an invaluable tool for genome annotation and the interpretation of proteomics data. This review is therefore aimed at providing a comprehensive overview on protein localization in *P. gingivalis* making use of an in-depth bioinformatic reappraisal of previously published biochemical, genomic, and proteomic studies (51, 59-85).

General architecture and subcellular compartments

To predict the subcellular localization of proteins in a bacterium, it is necessary to first gather information on this bacterium’s cellular architecture. The knowledge of subcellular compartments is in fact required to develop a species-specific prediction strategy. In this review, protein localization in a total of seven *P. gingivalis* strains was evaluated, including three reference strains and four clinical isolates. The three investigated reference strains W83, TDC60, and ATCC33277 are the main and best-studied *P. gingivalis* strains with publicly available genome sequences that were manually curated. Their proteomes were accessed and downloaded from UniprotKB (86) on 11th March 2018: W83 [UP000000588], ATCC 33277 [UP000008842], and TDC60 [UP000009221]. The included clinical isolates (20655, MDS140, MDS33, 512915) (46, 52) can be divided in PPAD sorting types I and type II (52). This classification concerns the differential sorting of PPAD as recently detected in one of our studies on clinical isolates (52). Compared to sorting type I isolates, sorting type II isolates display an extremely hampered production of OM- and OMV-bound PPAD, which appears to be due to a Gln to Lys amino acid substitution at position 373 of this protein (52). The two sorting type I isolates, 20655 and MDS140, were obtained from a patient with severe periodontitis but no RA, and a healthy carrier, respectively (46). The sorting type II isolates (512915 and MDS33), on the other hand, were isolated from a periodontitis patient without RA and a patient with severe periodontitis and RA, respectively (52). Of note, the previous study during which the sorting type I and II isolates were identified showed that neither the

association of PPAD with vesicles, nor the vesiculation of *P. gingivalis* by itself, are critical determinants for interactions of *P. gingivalis* with its human host (52), because the sorting type I or II distinction could not be reconciled with the severity of periodontitis according to the Dutch periodontal screening index (25).

Consistent with its status as a Gram-negative bacterium, the protein-containing subcellular compartments of a *P. gingivalis* cell can be divided in cytoplasm, inner membrane (IM), periplasm, and OM (Fig. 2). Nascent bacteriophages could, in principle, represent a separate intracellular compartment, but to date no bacteriophages have been described for *P. gingivalis* (87, 88). In addition, some of the proteins like PPAD are targeted to the extracellular milieu, in particular the periodontium of the human host (120). As mentioned above, proteins can be secreted either in a soluble state or bound to OMVs (Fig. 2) (47, 52, 89, 90). Gram-negative bacteria produce OMVs by natural "blebbings" of their outer membrane. Accordingly, OMVs consist of a single membrane originating from the OM, contain OM proteins, lipopolysaccharide (LPS), and other lipids. The cargo of OMVs also includes cytoplasmic and periplasmic proteins (91), but they appear enriched in virulence factors (89). The OMVs of *P. gingivalis* were shown to be involved in biofilm formation and to have invasive capabilities (90, 92, 93). In particular, the *P. gingivalis* OMVs were shown to enter host epithelial cells and degrade key receptor proteins using the afore-mentioned gingipains (94-96). These cysteine proteases are essential for virulence in animal models where they were shown to degrade many host proteins, thereby impairing cellular functions and the host immune response (97, 98). Other studies have implicated OMVs of *P. gingivalis* in selective Tumor Necrosis Factor tolerance (161), inflammasome activation and pyroptosis in macrophages (162). Therefore, OMVs represent non-viable satellite compartments of the *P. gingivalis* cell (Figs. 2 and 3). Of note, the precise mechanisms underlying the formation of OMVs in *P. gingivalis* are still unknown and, therefore, no bioinformatic tool exists or can be created yet to predict which proteins are localized in this peculiar extracellular compartment. While biochemical studies have investigated

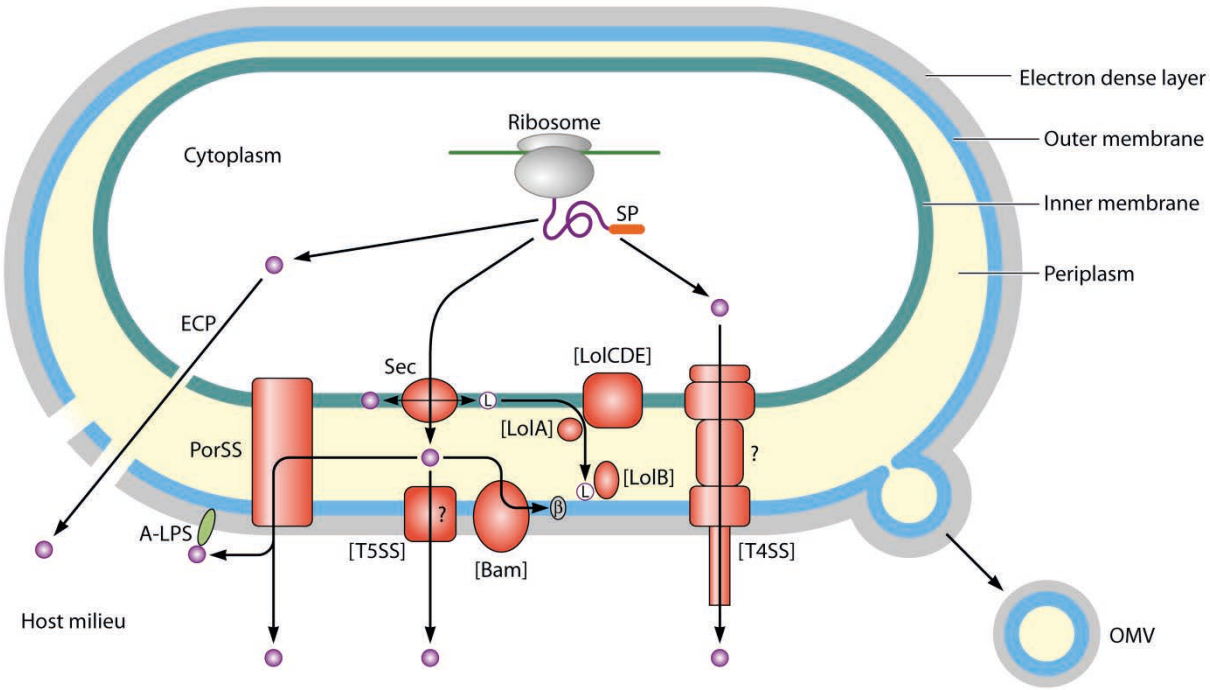


Figure 2. Protein sorting mechanisms in *P. gingivalis*. Overview of the cellular architecture and protein transport systems occurring in *P. gingivalis*. The indicated protein transport systems were identified by domain searches for major components of known transport systems previously identified in Gram-negative bacteria. The Lol, Bam, T4SS and T5SS systems for which only a limited number of known potential components were identified in *P. gingivalis* are indicated in parenthesis. SP, signal peptide.

the OMV cargo proteins (89), the results are presently still limited to the one strain analyzed. For this reason, the OMV compartment has not been taken into account in our bioinformatic reappraisal of the available data.

Systems for protein export from the cytoplasm, membrane insertion and secretion in P. gingivalis

Knowledge of the subcellular compartments present in *P. gingivalis* is required for the identification of protein transport, secretion, and membrane insertion systems. Uncovering the suite of such systems used by different *P. gingivalis* strains will grant a deeper understanding of the ways in which general virulence factors and particular toxins are exported from the cytoplasm and delivered to cells and tissues of the host. This will also highlight possible strain-specific differences. Importantly, certain transport systems are dedicated to the export of proteins with specific functions in virulence, but such systems can also serve other functions. The latter is showcased by the type IV protein secretion system that can also facilitate conjugation (99), or the type IX secretion system that is also employed in gliding motility (100). Therefore, a careful dissection of the occurrence of different types of protein sorting and secretion systems may lead to a detailed understanding of a strain's array of biological capabilities and virulence potential.

In general, Gram-negative bacteria possess an IM and an OM and, for this reason, several export and membrane insertion systems are utilized to translocate proteins across these two membranes, and to

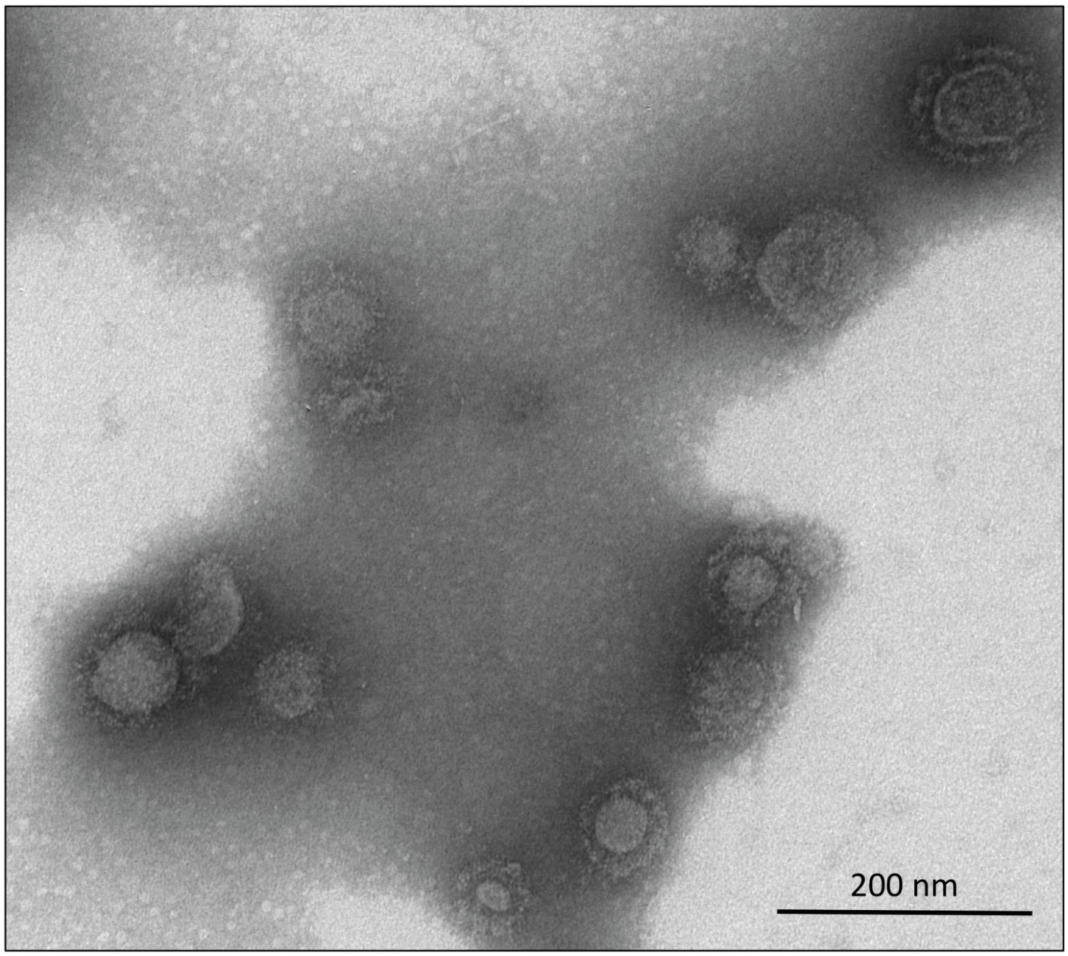


Figure 3. OMVs are the 'satellite compartments' of *P. gingivalis.* Transmission electron micrograph of purified outer membrane vesicles of the *P. gingivalis* type strain W83.

sort them to their rightful destinations. The vast majority of extracytoplasmic proteins in Gram-negative bacteria is translocated across the cytoplasmic membrane in an unfolded state by the Sec translocase. This includes integral IM proteins and lipoproteins that remain associated with the IM (101) (Fig. 2). A relatively small number of proteins traverses the cytoplasmic membrane *via* the Tat system, which is specific for cargo proteins in a pre-folded state that usually contain co-factors (102-104, 163). Among the proteins that reach the periplasm, the b-barrel proteins can be inserted into the OM by the 'b-barrel assembly machinery' (BAM) complex (105, 106), while lipoproteins are inserted into the OM by the 'localization of lipoprotein' (Lol) system (107, 108) (Fig. 2). Due to their major roles in protein sorting, these systems are broadly conserved among Gram-negative bacteria and their genes are, thus, easily recognizable by automated pipelines. Key components of the Sec, Tat, BAM, and Lol systems can be promptly identified by looking for the homologues of known members of these systems in other species. In addition, specific domain searches can be utilized (Table 1).

With the exceptions of SecE and SecG, either missing among some clinical strains or poorly annotated, all the analyzed strains of *P. gingivalis* possess the components of the SecYEG-DFyajC system. Intriguingly, however, they lack the known Tat translocase components, showing that the Tat system is not conserved in this bacterium. This is consistent with the outcome of domain searches using motifs identifying Tat signals (Tigr01409, Tigr01412, pfam10518), which yield no matches (Table 1), and with previous analyses reported in the literature (61). Moreover, only two members of the BAM system (BamA and BamC) appear to be present in the strains studied, as shown by domain searches and similarity analyses (Table 1). No homologues of BamB, BamD, and BamE are present in *P. gingivalis*. This is noteworthy, because only BamA and BamD are regarded universally essential for functionality of the Bam system (109). Similarly, the Lol system is only partially represented in *P. gingivalis* as merely four proteins with a LolE motif (COG4591) are detectable in the *P. gingivalis* reference strains. Some of these proteins display moderate levels of similarity to LolE proteins from other Gram-negative species, as judged by the presence of potential LolCE motifs (tigr002212, tigr002213). These proteins are predicted to reside in the IM, which would be consistent with the localization of the LolCE proteins of *Escherichia coli,* and half of them belong to the core proteome of *P. gingivalis.* Yet, canonical members of the Lol system, especially LolA, LolB, LolC, and LolD are absent from *P. gingivalis*. These observations suggest that analogous 'Lol' and 'BAM' systems may, respectively, be operational in the IM and OM of this bacterium (Fig. 2), while the prototype Lol and BAM systems are lacking.

Gram-negative bacteria can also possess other common systems enabling the translocation of proteins across the OM (110). These secretion systems vary from type I to type VIII (T1SS-T8SS), with the recent addition of a type IX secretion system specific to certain members of the Bacteroidetes phylum (61). The type IX secretion system is also referred to as T9SS, or Porin secretion system (PorSS), and the latter designation is most frequently used in the context of protein export in *P. gingivalis*. Unfortunately, secretion systems are usually not well annotated by automated pipelines, mainly because certain members of different secretion systems (e.g. T2SS and T4SS) share higher sequence similarity with one another than with functionally equivalent members of the same secretion system (e.g. pilin proteins). Moreover, many secretion systems are still poorly characterized, leading to difficulties in finding the most suited domains for a domain search. Fortunately, the genes encoding members of these systems usually co-localize on the genome, thus facilitating the identification of system components.

The potential presence of known secretion systems in *P. gingivalis* was evaluated *via* domain searches, literature and genome context analyses, and similarity searches across the *P. gingivalis* reference strains. All three analyzed reference strains lack the vast majority of secretion systems commonly encountered in Gram-negative bacteria (Table 1). Nevertheless, proteins containing two motifs belonging to members of the type I secretion system, pfam02321 and pfam03412, were found. The pfam02321 motif was detected in multiple proteins across all strains while pfam03412 was present in only two proteins for the TDC60 strain. Interestingly,

Table 1. Presence or absence of known protein transport and membrane insertion systems in *P. gingivalis.* Key members of protein transport systems and membrane protein insertion systems in the three *P. gingivalis* reference strains were identified by domain searches and secondary verification of the presence of particular orthologues. * Not annotated as protein.

SS	DOMAIN	PROTEIN	ATCC 33277	W83	TDC60
Sec	COG0653	SecA	PGN_1458	PG0514	PGTDC60_1633
	Tigr00963	SecA	PGN_1458	PG0514	PGTDC60_1633
	IPR003708	SecB			PGTDC60_1688
	IPR035958	SecB			
	IPR027398	SecD first TM region			
	IPR005791	SecD	PGN_1702	PG1762	PGTDC60_1374
	IPR005665	SecF	PGN_1702	PG1762	PGTDC60_1374
	IPR022645	SecD/F			
	COG0690	SecE	PGN_1577	PRESENT*	PGTDC60_1503
	Tigr00964	SecE	PGN_1577	PRESENT*	PGTDC60_1503
	COG1314	SecG	PGN_0258	PG0144	PGTDC60_0422
	Tigr00810	SecG	PGN_0258	PG0144	PGTDC60_0422
	COG0201	SecY	PGN_1848	PG1918	PGTDC60_0188
	Tigr00967	SecY	PGN_1848	PG1918	PGTDC60_0188
	COG0706	YidC	PGN_1446	PG0526	PGTDC60_1645
	Tigr03592	YidC	PGN_1446	PG0526	PGTDC60_1645
	Tigr03593	YidC	PGN_1446	PG0526	PGTDC60_1645
	COG1862	YajC	PGN_1485	PG0485	PGTDC60_1601
	Tigr00739	YajC	PGN_1485	PG0485	PGTDC60_1601
Sec			**+**	**+**	**+**
TAT	COG0805	TatC			
	Tigr00945	TatC			
	pfam00902	TatC			
	COG1826	TatA/E			
	Tigr01411	TatA/E			
	Tigr01410	TatB			
	Tigr01409	Tat signal			
	Tigr01412	Tat signal			
	pfam10518	Tat signal			
TAT			**-**	**-**	**-**
SRP	IPR004780	Ffh	PGN_1205	PG1115	PGTDC60_1100
	IPR004390	FtsY	PGN_0264	PG0151	PGTDC60_0428
SRP			**+**	**+**	**+**
BAM	Tigr03303	BamA	PGN_0299	PG0191	PGTDC60_0462
	IPR023707	BamA			
	Tigr03300	BamB			
	IPR017687	BamB			
	IPR014524	BamC			
	Tigr03302	BamD	PGN_1354	PG1215	PGTDC60_1188

(continued)

BAM	IPR017689	BamD	PGN_1354	PG1215	PGTDC60_1188
	pfam06804	BamD			
	pfam04355	BamE			
	IPR026592	BamE			
BAM			±	±	±
LOL	Tigr00547	LolA			
	pfam03548	LolA			
	COG2834	LolA			
	Tigr00548	LolB			
	COG3017	LolB			
	pfam03550	LolB			
	Tigr02212	LolC			
	Tigr02211	LolD			
	Tigr02213	LolE			
	COG4591	LolE	PGN_0718	PG0682	PGTDC60_0845
			PGN_0719	PG0683	PGTDC60_1224
			PGN_1025	PG0922	PGTDC60_1807
			PGN_1387	PG1252	PGTDC60_1808
LOL			±	±	±
T1SS	Tigr01842	PrtD			
	IPR010128				
	Tigr01843	HlyD			
	IPR010129				
	Tigr01844	TolC			
	IPR010130				
	Tigr01846	HlyB			
	IPR010132				
	Tigr03375	LssB			
	IPR017750				
	pfam02321	outer membrane efflux protein	PGN_0444	PG0063	PGTDC60_0345
			PGN_0715	PG0094	PGTDC60_0374
			PGN_1432	PG0285	PGTDC60_0631
			PGN_1539	PG0538	PGTDC60_1397
	IPR003423		PGN_1679	PG0679	PGTDC60_1540
			PGN_2012	PG1667	PGTDC60_1656
			PGN_2041		PGTDC60_1804
	pfam03412	bacteriocin exporter family (Peptidase C39 family)			PGTDC60_1000
	IPR005074				PGTDC60_1973
T1SS			-	-	-
T2SS	COG1450	PulD			
	COG2804	PulE			
	COG1459	PulF			
	COG2165	PulG			
	IPR013545	PulG			

(continued)

T2SS	Tigr02517	type II secretion system protein D (GspD)			
	IPR013356				
T2bSS T2bSS	Tigr02519	pilus (MSHA type) biogenesis protein MshL			
	IPR013358				
	Tigr02515	type IV pilus secretin (or competence protein) PilQ			
	IPR013355				
	pfam07655	Secretin N-terminal domain			
	IPR011514				
	pfam07660	Secretin and TonB N terminus short domain			
	IPR011662				
T2a-cSS, T3aSS	pfam00263	Bacterial type II and III secretion system protein (secretin)			
	IPR004846				
	pfam03958	Bacterial type II/III secretion system short domain			
	IPR005644				
T2SS			-	-	-
T3SS	COG1157	FliI			
	IPR032463	FliI			
	COG1766	FliF			
	IPR000067	FliF			
	COG1886	FliN			
	IPR012826	FliN			
T2a-cSS, T3aSS	pfam00263	Bacterial type II and III secretion system protein (secretin)			
T2a-bSS, T3aSS	pfam03958	Bacterial type II/III secretion system short domain			
T3aSS	Tigr02516	type III secretion outer membrane pore, YscC/HrcC family			
	IPR003522				
T3bSS	pfam02107	Flagellar L-ring protein (FlgH)			
	IPR000527				
T3SS			-	-	-
T4SS	COG3838	VirB2			
	IPR007039	VirB2			
	COG3702	VirB3			
	IPR007792	VirB3			
	COG3451	VirB4	PGN_0065	PG1481	PGTDC60_1018
					PGTDC60_1993
	COG3704	VirB6			
	IPR007688	VirB6			
	COG3736	VirB8			
	IPR007430	VirB8	PGN_0062	PRESENT*	PGTDC60_1021
	IPR026264	Type IV secretion system protein VirB8/PtlE			
	COG3504	VirB9			

(continued)

T4SS	IPR014148	VirB9			
	COG2948	VirB10			
	IPR005498	VirB10			
	COG0630	VirB11			
	IPR014155	VirB11			
	COG3505	VirD4	PGN_0076	PG1490	PGTDC60_1006
			PGN_0579		PGTDC60_1984
	IPR003688	Type IV secretion system protein TraG/VirD4		PG1490	PGTDC60_1984
T4bSS	pfam03524	Conjugal transfer protein			
	IPR010258				
	Tigr02756	type-F conjugative transfer system secretin TraK			
	IPR014126				
	pfam06586	TraK protein			
	IPR010563				
T4SS			±	±	±
T5SS	COG3468	adhesin AidA			
	COG5295	autotransporter adhesin			
	COG5571	autotransporter β-barrel domain			
T5cSS	pfam03895	YadA-like C-terminal region			
	IPR005594				
T5aSS	pfam03797	autotransporter β domain			
	IPR005546	autotransporter β domain	PGN_0129	PG1823	PGTDC60_0070
			PGN_0178	PG2130	PGTDC60_1255
			PGN_1744	PG2168	PGTDC60_1292
T5dSS	pfam07244	Surface ag VNR domain (PlpD POTRA motif)	PGN_0299	PG0191	PGTDC60_0462
	IPR010827				
	pfam01103	Bacterial surface Ag domain (PlpD β-barrel domain)	PGN_0147	PG0980	PGTDC60_0900
	IPR000184		PGN_0973	PG2095	PGTDC60_1324
T5SS			-	-	-
T5dSS			±	±	±
T6SS	Tigr03345	type VI secretion ATPase, ClpV1 family			
	IPR017729				
	Tigr03347	type VI secretion protein, VC_A0111 family			
	IPR010732				
	Tigr03350	type VI secretion system OmpA/MotB family protein			
	IPR017733				
	Tigr03352	type VI secretion lipoprotein, VC_A0113 family			
	IPR017734				
	Tigr03353	type VI secretion protein, VC_A0114 family			
	IPR010263				
	Tigr03354	type VI secretion system FHA domain protein			
	IPR017735				

(continued)

T6SS	Tigr03355	type VI secretion protein, EvpB/VC_A0108 family			
	IPR010269				
	Tigr03358	type VI secretion protein, VC_A0107 family			
	IPR008312				
	Tigr03362	type VI secretion-associated protein, VC_A0119 family			
	IPR017739				
	Tigr03373	type VI secretion-associated protein, BMA_A0400 family			
	IPR017748				
T6SS			-	-	-
T7SS	Tigr03919	type VII secretion protein EccB			
	IPR007795				
	Tigr03920	type VII secretion integral membrane protein EccD			
	IPR006707				
	Tigr03921	type VII secretion-associated serine protease mycosin			
	IPR023834				
	Tigr03922	type VII secretion AAA-ATPase EccA			
	IPR023835				
	Tigr03923	type VII secretion protein EccE			
	IPR021368				
	Tigr03924	type VII secretion protein EccCa			
	IPR023836				
	Tigr03925	type VII secretion protein EccCb			
	IPR023837				
	Tigr03926	type VII secretion protein EssB			
	IPR018778				
	Tigr03927	type VII secretion protein EssA/YueC			
	IPR018920				
	pfam10661	WXG100 protein secretion system (Wss), EssA			
	IPR034026				
	Tigr03928	type VII secretion protein EssC			
	IPR023839				
	IPR022206				
	Tigr03931	type VII secretion-associated protein, Rv3446c family			
	IPR023840				
	pfam00577	Fimbrial usher protein			
	IPR000015				
	pfam06013	Proteins of 100 residues with WXG			
	IPR010310				
T7SS			-	-	-
T8SS (ENP)	pfam03783	Curli production assembly/ transport component CsgG			
	IPR005534				
	pfam07012	Curlin associated repeat			
	IPR009742				

(continued)

T8SS (ENP)	pfam10614	Tafi-CsgF			
	IPR018893				
	pfam10627	CsgE			
	IPR018900				
T8SS (ENP)			-	-	-
PorSS		PorK	PGN_1676	PG0288	PGTDC60_1400
		PorL	PGN_1675	PG0289	PGTDC60_1401
		PorM	PGN_1674	PG0290	PGTDC60_1402
		PorN	PGN_1673	PG0291	PGTDC60_1403
		PorP	PGN_1677	PG0287	PGTDC60_1399
		PorQ	PGN_0645	PG0602	PGTDC60_1728
	pfam13568	PorT	PGN_0778	PG0751	PGTDC60_1868
		PorU	PGN_0022	PG0026	PGTDC60_0023
		PorV (PG27,LptO)	PGN_0023	PG0027	PGTDC60_0024
		PorW	PGN_1877	PG1947	PGTDC60_0218
	pfam14349	Sov	PGN_0832	PG0809	PGTDC60_1927
		PorX	PGN_1019	PG0928	PGTDC60_0851
		PorY	PGN_2001	PG0052	PGTDC60_0334
		Lipoprotein; TPRd, WD40d, CRDd, OmpA family domain	PGN_1296	PG1058	PGTDC60_0980
		PorZ	PGN_0906	PG1064	PGTDC60_1144
	Orthology	PorZ	PGN_0509	PG1604	PGTDC60_0697
		β-barrel protein	PGN_0297	PG0189	PGTDC60_0460
		TonB-dependent receptor; β-barrel protein	PGN_1437	PG0534	PGTDC60_1652
		Omp17; OmpH-like	PGN_0300	PG0192	PGTDC60_0463
		sigP	PGN_0274	PG0162	PGTDC60_0438
PorSS			+	+	+

these two proteins display no significant similarity to proteins in the other *P. gingivalis* reference strains as opposed to a significant similarity shared with proteins belonging to other species in the same phylum. Of note, the pfam02321 motif can also detect OM components of drug and metal efflux pumps, suggesting that the identified proteins do not necessarily belong to a functional type I secretion system. Conversely, the pfam03412 motif was used in combination with tigr01193 to identify bacteriocin exporters. The two proteins identified in strain TDC60 appear to possess both motifs, suggesting a possible role in bacteriocin secretion. However, the similarity scores for the tigr01193 motifs are significantly lower than those for the pfam03412 motifs. In conclusion, it appears that a canonical type I secretion system is absent from *P. gingivalis* (61).

None of the known protein components of type II and III secretion systems was found in *P. gingivalis*, including members of subclasses a, b, and c of the type II secretion systems and subclasses a and b of the type III secretion systems. Conversely, domain searches for three major components of the type IV secretion system, named VirB4, VirB8, and VirD4, showed multiple matches across the three reference strains. The VirB4 domain is present in two TDC60 proteins that share a relatively high level of similarity, while the VirD4 domain is present in two W83 and two TDC60 proteins. At least one gene per strain encoding these proteins co-localizes with a VirB4 motif gene on the *P. gingivalis* chromosome, with a distance between the respective genes of about 5 kb in the W83 strain, about 7 kb in the ATCC 33277 strain, and 10-11 kb for the TDC60 strain.

The VirB8 domain is present in one gene per reference strain, but it should be noted that the respective gene has not been annotated for the W83 strain and that the presence of the VirB8 domain was only discovered upon closer inspection of the genome sequence.

Although no matches were identified for signature domains of key components of the type V secretion system, one protein in every reference strain was found to display a PlpD motif (pfam07244). This motif identifies components of subclass d of the type V secretion system. Moreover, these proteins appear to possess also a second PlpD motif used in T5dSS searches, pfam01103, although this second motif was identified with a sensibly lower score. On this basis, it is difficult to predict the activity of T4SS and T5dSS in the analyzed *P. gingivalis* strains. In the canonical T4SS, VirB4 and VirD4 are two of the three ATPases that energize the secretion machinery (111). This could imply that the VirB4- and VirD4-like proteins of *P. gingivalis* may be involved in another secretion system, or that they serve a different function. In contrast, the possibility that a T4SS could function in *P. gingivalis* in the absence of other key members of this type of secretion system appears less likely. Another piece of evidence in line with the latter view relates to the fact that the VirB8 domain, used for the present similarity searches, also recognizes conjugal transfer proteins, like TrbF and TraK. Nonetheless, VirB8 is generally responsible for forming the channel through which the T4SS cargo proteins are translocated across the IM. Hence, the detected *P. gingivalis* proteins containing a VirB8 domain could potentially offer an alternative pathway to the Sec system for protein passage across the IM of this bacterium (Fig. 2).

No proteins belonging to the type VI, VII, or VIII secretion systems were detectable in the analyzed *P. gingivalis* strains, which is in agreement with previous literature (61). On the other hand, *P. gingivalis* strongly relies on a novel secretion system shared by members of the Bacteroidetes phylum, the afore-mentioned PorSS (61), whose prominence in Bacteroidetes has recently been highlighted (1) (Fig. 2). The PorSS comprises several proteins broadly conserved throughout the Bacteroidetes group and is also involved in gliding motility in many species of this phylum, albeit not in *P. gingivalis*. As of now, there is general consensus that 17 proteins are essential for the PorSS function, although two additional proteins are probably required as well (84, 85). Four of these proteins, PorK-N, form the PorSS core membrane complex. PorM, the main component of this complex, appears to localize in the IM together with PorL. However, thanks to its long periplasmic domain, PorM is capable of interacting with the rest of the complex comprising the OM-bound lipoprotein PorK and the periplasmic OM-bound protein PorN (61, 112). Cargo proteins of the PorSS are targeted for secretion by conserved C-terminal domains (CTDs) (61), which can be identified by the TIGR04131 (IPR026341) and TIGR04183 (IPR026444) motifs. The presence of a Sec-type N-terminal signal peptide in proteins exported *via* the PorSS suggests that these proteins are translocated across the IM by the Sec machinery. Importantly, all known members of the PorSS are present in all of the strains evaluated here, highlighting the major role that this export system plays in *P. gingivalis* (Fig. 2).

It was demonstrated by different localization studies that certain CTD-containing proteins cross the OM with the help of the PorSS, subsequently appearing both in the OM and in the extracellular milieu (51, 113, 114). According to the current models, the CTD is cleaved during export of the 'CTD proteins' by a sortase-like mechanism, and the resulting mature proteins are secreted or re-attached to the OM *via* A-LPS modification, with the A-LPS acting as an anchor to the bacterial surface (84, 85). Consequently, the CTD is lacking from the mature soluble forms of these proteins (115, 116), and it has not been detected in the OM-associated mature forms, which are extensively A-LPS modified (117-119). Clearly, especially due to the two possible destinations of CTD proteins (i.e. OM insertion or secretion), it is challenging to predict their precise localization by bioinformatic approaches.

In addition to the classical secretion systems, the afore-mentioned release of OMVs with cargo proteins (Fig. 3) should be regarded as a specialized protein secretion pathway dedicated to virulence and the capture of nutrients (89). Indeed, the mechanism triggering the blebbing process that leads to these

nanostructures, albeit poorly understood, is not random (93). Additionally, the proteins secreted *via* this pathway seem to empanel mostly periplasmic and OM proteins, which serve as virulence factors (89, 120). Among the latter, proteases, especially the gingipains, appear to be most abundant (89). The prevalence of proteases in the OMVs might serve several purposes. Firstly, it could be a way to deliver them to their foreign targets, especially proteins of phagocytic cells (120). Secondly, encasing the proteases within the membrane of the OMVs can protect them and/or the rest of the OMV cargo from the outside environment, either physically or by rendering proteolytic sites on these proteins inaccessible. Thirdly, this feature might have evolved to protect *P. gingivalis* proteins, for example bound to the OM, from the bacterium's own highly proteolytic potential. Lastly, OMVs and the OMV-associated PPAD could serve a decoy function in immune evasion by *P. gingivalis* (120). In light of these different scenarios, the observed phenomenon of extracellular compartmentalization through vesiculation might be categorized as a 'protective secretion' behavior. In fact, the attachment of CTD proteins, like PPAD, to the OM and to OMVs *via* A-LPS modification seems to protect them from proteolysis by *P. gingivalis*' own proteases, as evidenced by the recent observation that OMV disruption by ultrasound results in PPAD degradation (121). In addition, the finding that the PPAD proteins of sorting type II isolates, which are ineffectively attached to the OM and OMVs, are processed to a 37 kDa form is consistent with the idea that the OMVs serve to protect cargo against proteolysis (52).

Signal peptidases

The identification of the suite of secretion systems in *P. gingivalis* warranted a further investigation on the signal peptidases involved. Firstly, the Sec system utilizes two different types of signal peptidases. In general, cargo proteins of the Sec pathway are processed by signal peptidase I, which belongs to the S26 Merops family (122-124). As is the case for all living cells, the *P. gingivalis* genome encodes signal peptidase I, as identified through COG0681 and pfam00717 searches. Of note, a recent study from Bochtler *et al*. showed that over 60% of signal peptidase I substrates in *P. gingivalis* display a glutamine residue immediately downstream of the signal peptidase I cleavage site (in position +1), irrespective of their subcellular localizations (125). These glutamine residues are cyclized to pyroglutamate residues by the glutaminyl cyclase PG2157 (alternatively called PG_RS09565), a lipoprotein most likely located in the IM (125). This high frequency of signal peptidase I substrates with a glutamine residue in position +1 is a common feature of most Bacteroidetes species (125). Lipoproteins have N-terminal signal peptides recognized and removed by the signal peptidase II, which takes place after the invariant cysteine residue at position +1 relative to the cleavage site has been diacyl-glyceryl modified by the diacyl-glyceryl transferase Lgt (123). Signal peptidase II belongs to the A08 Merops family and is detectable in *P. gingivalis* through domain searches for the pfam01252 and COG0597 motifs. Likewise, Lgt is conserved in all investigated *P. gingivalis* strains as confirmed by BLAST searches. In *E. coli*, the N-terminal amino group of the diacyl-glyceryl-modified cysteine of the mature lipoprotein is acylated by the N-acyl transferase Lnt (123). This may not be the case in *P. gingivalis*, as no homologues of the *E. coli* Lnt were detected in the investigated strains. However, the possibility of N-acylation of the mature lipoprotein upon cleavage by Lsp cannot be fully excluded, since it was shown that N-acylation by an as yet unidentified enzyme takes place in *Staphylococcus aureus* (126). Interestingly, the N-acylation of staphylococcal lipoproteins has been invoked in the silencing of innate and adaptive immune responses (126), which is a trait that could enhance the fitness and pathogenicity of *P. gingivalis* as well.

Available algorithms for genome-wide identification of exported bacterial proteins

Genome-wide prediction of the subcellular localization of proteins is a relatively recent endeavor in proteomics that has garnered increasing attention, because it provides valuable insights into the biological functions of the sorted proteins, even if their precise function is still unknown (127-130). Various bioinformatic tools have been

designed to identify signal peptides, such as SignalP (131), Predisi (132) and Phobius (133). These algorithms are generally used to predict signal peptides cleaved by signal peptidase I, but they do not readily recognize the lipoprotein signal peptides that are cleaved by signal peptidase II. To address this issue, lipoproteins have to be identified first by predictors capable of recognizing lipoprotein signals, such as LipoP (134) and Lipo (135). The subsequently developed PSORT I represented the first comprehensive bacterial protein localization predictor. Since then, several prediction tools for protein localization have been developed and implemented, rendering bioinformatic approaches a viable alternative to biochemical localization studies (130, 136, 137). All these studies involved the development of a complex network of subcellular localization predictors that were tailored to a specific bacterium in order to predict, as accurately as possible, the position of each protein in the proteome. One of these studies (137) has been taken into particular consideration for this review, and its workflow was adapted to review the overall protein localization in *P. gingivalis*.

It should be noted that all publicly available prediction tools for subcellular protein localization have particular pro's and con's. One of the difficulties in selecting the most suited programs for a bacterium of interest lies in the fact that publicly available predictors may quickly cease to be maintained, are subject to major modifications, or even become obsolete. This, coupled with the fact that certain programs may be more suited to bacteria of a certain group, makes it difficult to implement strategies previously developed for major model organisms, such as *E. coli* or *Bacillus subtilis* (123, 127, 129, 138). Aside from public access, another important parameter determining our choice of programs was availability of a batch submission option, which grants fast genome-wide analyses. Moreover, to further refine the selection of prediction programs for a comprehensive overview of subcellular protein localization in *P. gingivalis*, tools with a high level of specialization were used as listed in Table 2. In most cases, such tools were single function predictors with few limitations, especially limitations that could have been offset by the application of other programs.

Interestingly, different predictors occasionally assigned the same proteins to different subcellular compartments, even in case of programs with the same specific functions. Disagreements in localization between different programs underscore the notion that some predictors may be more accurate or, at the very least, better suited than others to chart the proteins of a specific bacterium. Moreover, these discrepancies reveal the levels of uncertainty of bioinformatics predictions and the need for an organized method encompassing all the chosen tools that can exploit all the strengths and balance the limitations of each program. On the other hand, it must be acknowledged that protein sorting mechanisms in a living bacterium do not usually operate with a fidelity of 100%, which means that proteins that are generally secreted are detectable within different cellular compartments, while proteins that are meant to be retained in the cell (e.g. cytoplasmic proteins, lipoproteins, or cell wall-bound proteins) can be encountered in the extracellular environment. The protein sorting ambiguities encountered *in silico* are thus perhaps an unintended reflection of the imperfections of sorting systems employed by a bacterial cell *in vivo*. Clearly, as long as these imperfections have no bearing on the competitive success of a bacterium, they do not matter.

To meet the need for biologically relevant predictions of protein sorting, a decision tree (Fig. 4) was devised, which organizes the predictors and sorts proteins through them with the purpose of assigning them to their rightful subcellular compartment. The first challenge in a prediction analysis is to localize the components of the export, secretion, and membrane insertion systems themselves, which relates to the difficulty in recognizing their signal peptides by predictors. The level of difficulty depends on the system examined, with more common and conserved systems being more easily localized. For example, some components of the recently discovered Por secretion system have an uncertain localization. Secondly, the identification of lipoproteins has priority, especially in view of the inability of different predictors to distinguish Sec signal peptides cleaved by the lipoprotein-specific signal peptidase II from Sec signal peptides cleaved by signal peptidase I. Notably, localization tools generally distinguish between IM and OM lipoproteins, utilizing data from extensive research on the widely favored model Gram-negative bacterium *E. coli*. These studies

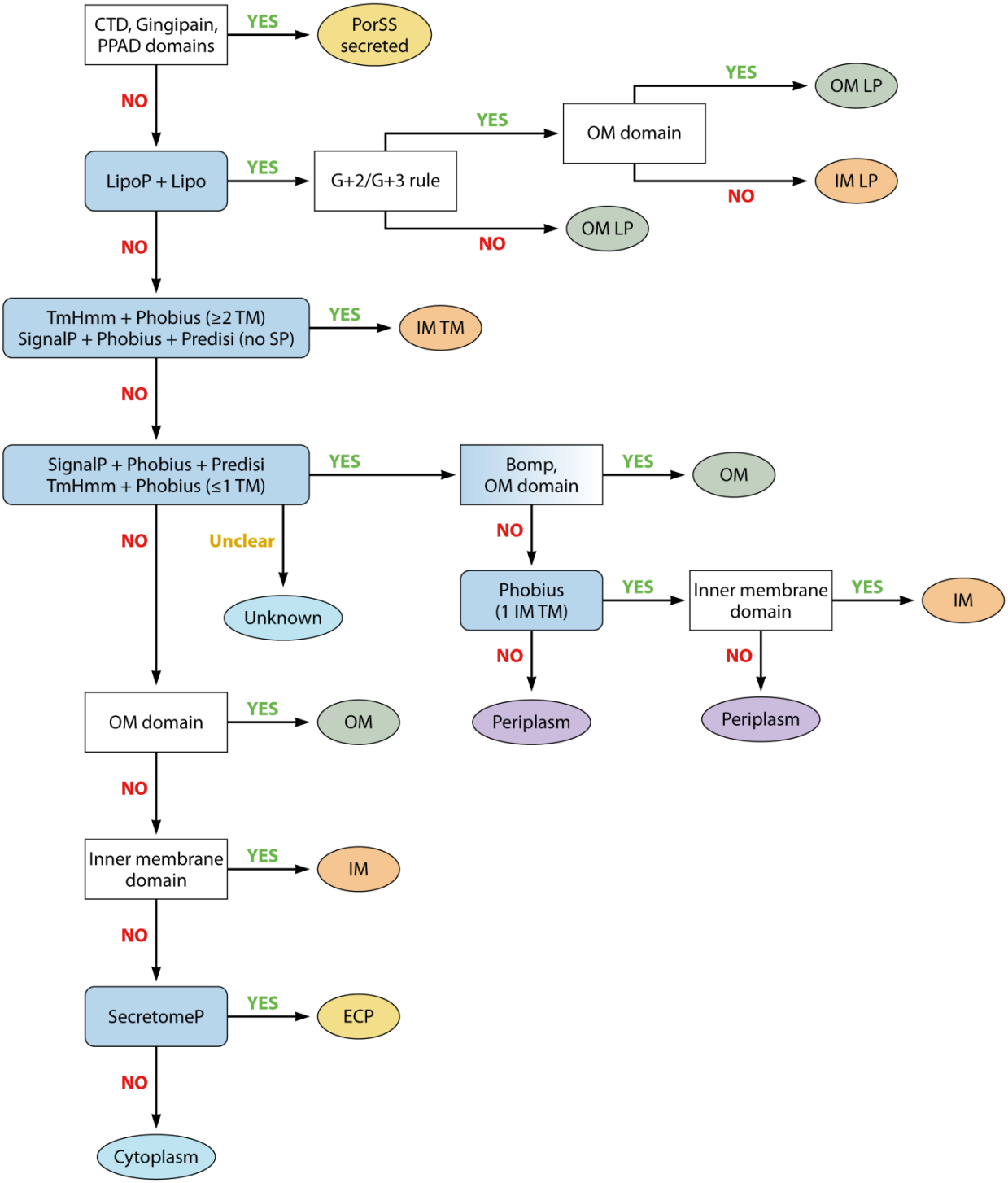


Figure 4. Bioinformatics pipeline to unravel protein sorting events in *P. gingivalis*. The flowchart depicts the different steps employed to assess the subcellular localization of proteins in the analyzed *P. gingivalis* strains.

Table 2. List of localization predictors. Overview of localization predictors, membrane insertion detectors, and other programs used in this study and their relative strengths and weaknesses.

NAME	USE	LIMITATIONS
LipoP	primarily prediction of Sec signal peptides that are cleaved by SpII but also provides prediction of inner membrane or cytoplasmic localization as well as SpI cleavage	does not detect Tat substrates
Lipo	prediction of Sec signal peptides cleaved by SpII	does not detect Tat substrates
SignalP	prediction of Sec signal peptides cleaved by SpI	does not detect Tat substrates
Predisi	prediction of Sec signal peptides cleaved by SpI	does not detect Tat substrates
Phobius	prediction of alpha helices in inner membrane proteins, distinguishing N-terminal TM from signal peptides	
TmHmm	prediction of alpha helices in inner membrane proteins	signal peptides often considered TM spans
Bomp	prediction of beta-barrel spans in outer membrane proteins	
SecretomeP	prediction of ECP	limited number of sequence per batch
Interpro	functional analysis of proteins by classification into families, domain and site prediction by combination of protein signatures from a number of member databases	

have shown that lipoproteins possessing an aspartic acid in the +2 position of the mature protein become IM lipoproteins (i.e. the 'D^{+2} rule'), while all other lipoproteins are presented to the OM by the Lol system (139). Intriguingly, several exceptions to this rule have been observed in other species (125, 138, 140-143), presenting the possibility that it is only obeyed in *Enterobacteriaceae*. Analyzing known OM lipoproteins of *P. gingivalis* by applying the D^{+2} rule, in fact, resulted in a faulty prediction for the subcellular location of the vast majority of lipoproteins. Conversely, inspection of lipoproteins of known subcellular localization showed a preferential glycine residue at the +2 or +3 positions of the mature form for IM lipoproteins. The present evaluation of lipoprotein localization in *P. gingivalis* therefore relied on inspection of the +2 and +3 residues combined with specific domain searches. Following the designation of the 'lipoproteome', investigation of proteins with transmembrane helices and Sec signal peptides was performed, in this order (Fig. 4). This relates to the fact that predictors of membrane spanning regions occasionally mistake relatively longer signal peptides for transmembrane spans (Table 2).

Excretion of cytoplasmic proteins (ECP), also termed non-classical or leaderless secretion, is a highly discussed topic and a way to explain the presence in the extracellular milieu of proteins that lack a known signal peptide and a dedicated transport system of the categories described above (144, 145). These features apply to the bulk of cytoplasmic proteins. Accordingly, cell lysis was for a long time the most accredited hypothesis to explain the presence of cytoplasmic proteins in the extracellular milieu (146). This view is supported by the observation that ECP can be associated with autolysin and phage activity, or the production of cytotoxic peptides (145, 164). Nonetheless, the existence of dedicated 'non-classical' secretion systems for proteins deprived of known signal peptides cannot be excluded, as underpinned by the relatively recent

discovery of the Tat and type VII secretion systems (147). Such hidden treasures are likely to be buried in the exoproteome haystack, until uncovered by the application of molecular biological or mass spectrometric approaches to assess bacterial protein secretion. In fact, with increasing sensitivity of mass spectrometric measurements, more and more signal peptide-less proteins have been identified in bacterial exoproteomes. This is exemplified by a recent investigation on the exoproteome of *P. gingivalis*, where many signal peptide-less proteins were identified in the growth medium fraction (53). In fact, the latter analysis highlights two remarkable features. Firstly, signal peptide-less extracellular proteins were overrepresented amongst the low-abundance extracellular proteins and, secondly, the detection of these proteins was most variable between the investigated strains. This is suggestive of an unspecific export mechanism, such as cell lysis. Yet, also amongst the most abundantly detectable and invariant exoproteins of *P. gingivalis* there are proteins lacking signal peptides, which is suggestive of specific export, stable extracellular maintenance in the presence of gingipains, and a possible function in the bacterial life cycle. As to possible functions, it has been shown that proteins with important roles in the cytoplasm, like elongation factors and proteins involved in central carbon metabolism, can serve important extracytoplasmic 'moonlighting' functions in bacterial adhesion to mammalian cells and tissues (145, 148, 149). Altogether, it seems that ECP in Gram-negative bacteria may be more complex than initially thought, with several distinct pathways present (144). Importantly, proteins subject to ECP can be predicted by homology using SecretomeP 2.0 (150).

Dedicated pipeline to approximate subcellular protein localization in *P. gingivalis*

To approximate subcellular protein localization in *P. gingivalis* with the ultimate objective of better understanding which proteins are targeted to the bacterial cell envelope or the host milieu, an in-house script implementing the decision tree presented in Figure 4 was developed. In addition to the afore-mentioned algorithms, TMHMM (151) was included to predict transmembrane helices, BOMP (152) to predict β-barrel OM proteins, and InterPro Scan (153) version 5.27 to detect particular domains in the InterPro consortium database (154) version 66. Based on *P. gingivalis* proteins of known location (Table S1), three lists of domains specific for PorSS cargo, IM proteins, and OM proteins were established (Table S2). Such mainly structural domains were chosen to be as specific as possible, in order to avoid biases. Using the software listed in Table 2, localization data were generated for all the seven revisited *P. gingivalis* strains. Further, following the flow scheme presented in Figure 4, a knowledge-based approach was implemented that is grounded on the currently available understanding of protein sorting systems active in *P. gingivalis*, as detailed in the afore-mentioned sections. Importantly, the hierarchy of decisions in this pipeline was tailored to minimize mistakes and biases, and to maximize compensation for possible software weaknesses.

Proteins displaying at least one of the selected PorSS cargo-specific domains (Table S2) were immediately designated as secreted *via* the PorSS (Fig. 4), as these signatures are highly reliable in predicting secretion *via* this pathway. On this basis, the inspected *P. gingivalis* strains potentially secrete between 19 and 24 proteins specifically via the PorSS (Table 3). Despite its high specificity, this approach does not guarantee the identification of all PorSS cargo proteins, because some proteins exported *via* the PorSS may lack the selected PorSS cargo domains. Further, the present listing of predicted PorSS cargo proteins (Table S3) may represent an underestimation since potential misidentifications were not manually curated in order to avoid bias. This explains why the present number of potential PorSS cargo proteins is lower than the previously proposed 30 to 35 cargo proteins (85, 155), which may include some proteins whose secretion is indirectly related to the PorSS.

In the second step of the prediction pipeline, both the LipoP and Lipo algorithms were used to identify lipoproteins amongst those proteins that were not assigned as PorSS cargo (Fig. 4). Since there was

Table 3. Summary of predicted protein localizations. Overview of the protein localization predictions for each strain enumerating all the proteins present in the different subcellular compartments. CYT = cytosolic protein; ECP = ECP protein; IM = inner membrane protein; IM LP = inner membrane lipoprotein; OM = outer membrane protein; OM LP = outer membrane lipoprotein; PERI = periplasmic protein; PorSS = PorSS secreted protein; UNK = protein of unknown localization; TOT EXTRA = total of extracellular proteins; TOT IM = total of inner membrane proteins; TOT OM = total of outer membrane proteins.

ATCC 33277		**W83**		**TDC60**		**MDS33**		**MDS140**		**512915**		**20655**	
CYT	1286	CYT	1193	CYT	1451	CYT	1375	CYT	1362	CYT	1426	CYT	1366
ECP	95	ECP	80	ECP	109	ECP	94	ECP	97	ECP	103	ECP	115
IM	3	IM	3	IM	3	IM	3	IM	3	IM	3	IM	3
IM LP	18	IM LP	19	IM LP	18	IM LP	20	IM LP	22	IM LP	20	IM LP	19
IM TM	333	IM TM	316	IM TM	348	IM TM	332	IM TM	318	IM TM	342	IM TM	363
OM	57	OM	58	OM	60	OM	61	OM	56	OM	62	OM	64
OM LP	67	OM LP	46	OM LP	54	OM LP	54	OM LP	55	OM LP	61	OM LP	56
PERI	135	PERI	118	PERI	118	PERI	138	PERI	125	PERI	131	PERI	136
PorSS	22	PorSS	22	PorSS	24	PorSS	20	PorSS	21	PorSS	19	PorSS	20
UNK	6	UNK	8	UNK	9	UNK	9	UNK	6	UNK	5	UNK	5
Total	**2022**	**Total**	**1863**	**Total**	**2194**	**Total**	**2106**	**Total**	**2065**	**Total**	**2172**	**Total**	**2147**
TOT EXTRA (ECP + PorSS)	117	**TOT EXTRA** (ECP + PorSS)	102	**TOT EXTRA** (ECP + PorSS)	133	**TOT EXTRA** (ECP + PorSS)	114	**TOT EXTRA** (ECP + PorSS)	118	**TOT EXTRA** (ECP + PorSS)	122	**TOT EXTRA** (ECP + PorSS)	135
TOT IM (IM + IM LP + IM TM)	354	**TOT IM** (IM + IM LP + IM TM)	338	**TOT IM** (IM + IM LP + IM TM)	369	**TOT IM** (IM + IM LP + IM TM)	355	**TOT IM** (IM + IM LP + IM TM)	343	**TOT IM** (IM + IM LP + IM TM)	365	**TOT IM** (IM + IM LP + IM TM)	385
TOT OM (OM + OM LP)	124	**TOT OM** (OM + OM LP)	104	**TOT OM** (OM + OM LP)	114	**TOT OM** (OM + OM LP)	115	**TOT OM** (OM + OM LP)	111	**TOT OM** (OM + OM LP)	123	**TOT OM** (OM + OM LP)	120

no possibility for a majority vote, the LipoP predictions were given priority in case of disagreement. The same approach was used to assess the signal peptidase II cleavage sites, which was necessary to pinpoint amino acid residues at positions +2 and +3 of the mature lipoproteins. If glycine residues were absent from these positions, the respective protein was predicted to be an OM lipoprotein (OM LP). If a glycine residue was present at the +2 or +3 position, an additional control was performed by assessing the presence of a known OM domain. If an OM domain was detected (Table S2), the '$G^{+2/+3}$ rule' was ignored and the protein was still predicted as an OM lipoprotein (OM LP). Conversely, the apparent lack of OM domains resulted in a protein's designation as an IM lipoprotein (IM LP). Per investigated strain, the numbers of predicted IM lipoproteins ranged from 18 to 22, and the numbers of OM lipoproteins from 46 to 67 (Table 3). The relatively large variation in the numbers of OM lipoproteins predicted for different strains may relate to previously observed genomic rearrangements in *P. gingivalis* (156).

For non-lipoproteins, an agreed number of two or more transmembrane helices as identified by TMHMM and Phobius was used to predict IM transmembrane proteins (Fig. 4). When instead both programs agreed on the presence of at least one transmembrane helix, a signal peptide check was performed, in order

to reduce the number of false-positively predicted helices caused by the presence of a signal peptide. When the signal peptide prediction consensus was one or lower (i.e. one or none of the three applied programs predicted a signal peptide) the predicted helix was considered a 'true positive', and the respective protein was therefore predicted to have a transmembrane localization in the IM. In case of a signal peptide consensus equal to, or higher than two (i.e. at least two out of the three applied programs predicted a signal peptide) and an agreed number of transmembrane helices as predicted by TMHMM and Phobius was equal or lower than one, the protein was considered to have a signal peptide. In this case, it was further analyzed to determine a possible IM, OM, or periplasmic localization. Conversely, when the signal peptide consensus was one or zero, and the agreed number of transmembrane helices as predicted by TMHMM and Phobius was one or zero, the protein was considered to lack a signal peptide. Thus, despite being allegedly unable to cross the IM *via* Sec, such a protein was further analyzed for a possible cytoplasmic localization, or a potential IM, OM, or extracellular localization *via* ECP. In all other cases, i.e. when the outcome of the predictions of transmembrane helices and the presence of a signal peptide were in conflict, the predicted localization of the respective protein was designated as unknown (Table 3). Merely five to nine proteins with unknown localization were encountered for the presently evaluated strains, suggesting that the adopted approach was extremely discriminative and robust.

Proteins with a signal peptide are able to cross the IM, ending up in the periplasm. The additional presence of either a β-barrel, indicated by a BOMP score higher than three, or of at least one OM domain (Table S2), can be considered as an indicator for subsequent association with or insertion into the OM. The latter proteins were thus predicted to localize in the OM compartment. In case a transmembrane domain, β-barrel, or OM domain were absent, while a signal peptide was present, the respective protein was designated as having a periplasmic localization. Since the presence of a Phobius-predicted transmembrane domain is indicative of protein retention in the IM, such proteins were designated as IM resident-proteins.

In the absence of a canonical signal peptide, a protein can be retained in the cytosol, be secreted through non-canonical or unknown pathways thereby ending up in the OM or extracellular milieu, or be inserted into or associated with the IM. For such reasons, all proteins lacking a predicted signal peptide were checked for the presence of OM domains (Table S2). If one or more of such OM domains were present, the respective protein was designated to have an OM localization. Analogously, the presence of an IM-related domain (Table S2) was used as an indicator for IM localization.

A SecretomeP analysis was performed as the last verification step in the prediction pipeline (Fig. 4), because of its knowledge-based nature. At this juncture, the remaining proteins exhibit no relevant feature as discussed above and, accordingly, their predicted sorting destination could only be the cytoplasm or the extracellular milieu due to non-canonical or unknown ECP pathways. Therefore, in the presence of a SecretomeP score equal or higher than 0.75, proteins were predicted to undergo ECP. Instead, a lower score pointed at a cytosolic localization, since none of the applied predictors suggested the possibility of the respective protein leaving the cytosol. The overall outcome of the predicted protein localization in *P. gingivalis* is listed in Table S3, while Table 3 presents an overview of these predictions.

Core and variant exoproteome analyses

Interestingly, analysis of the *P. gingivalis* exoproteome highlighted strain-specific variations (53), which were also encountered in the present inspection of subcellular protein localization. This was the incentive for a bioinformatics-based appraisal of the core and variant (exo)proteome of *P. gingivalis*. Thus, to identify orthologues in the proteome of different strains, reciprocal best hits (RBHs) were calculated. In brief, Galaxy (157) was used to perform reciprocal protein BLAST searches (NCBI BLAST+ v. 2.3.0 (158)). Default parameters (minimum percentage identity: 70%; minimum High Scoring Pair (HSP) coverage: 50%) were used and all redundancies were removed prior to the BLAST search. RBHs were then calculated by blasting the deduced

amino acid sequences of all investigated strains against those of *P. gingivalis* ATCC 33277. Despite *P. gingivalis* W83 being the most used reference strain in the field, the ATCC 33277 strain was adopted as a reference for the present analyses after the realization that many proteins were actually encoded by the W83 genome sequence while the respective genes were never annotated (data not shown). Tblastn was used to identify some of these proteins, being part of the main secretion complexes, and they are reported in Table 1. The core proteome was thus defined by the set of proteins having an ortholog in all six strains analyzed against the ATCC 33277 reference strain (Table S4). The remaining protein complement identified for each strain is regarded as the respective variable proteome (Table S5). Of note, considering possible misannotations of the used genome sequences, the presently proposed distinction between the *P. gingivalis* core and variable proteomes should be regarded as an approximation rather than an absolute distinction.

To predict the core exoproteome, the proteins in the core proteome were divided according to their possible subcellular localizations, as per our prediction pipeline, and two categories were pulled together: 1) proteins of the OM compartment (OM_LP and OM proteins) and 2) PorSS cargo proteins (Table S6). The GO terms associated with the domains detected by InterPro for these exoproteins were taken into account for each strain. The obtained GO terms were then used in a REVIGO (159) analysis, to unravel the network of biological pathways created by the core exoproteome. It should be noted that the potential ECP complement as designated by our pipeline was excluded from the exoproteome classification due to its high variability between strains (Table S3). The remaining predicted exoproteins were, instead, almost entirely predicted to make up the core exoproteome. The limited suitability of SecretomeP for our *P. gingivalis* dataset is probably due to the fact that ECP predictors cannot be tailored on specific transport systems or bacterial species due to their intrinsic nature. GO term analysis of the core exoproteome predicted for each inspected strain yielded close to identical results (Fig. 5) with some marginal differences observed for strain MDS33 (data not shown). The latter may relate to minor discrepancies in the genome annotation of strain MDS33, or to some small potential differences in the core orthologs of this strain. The identified core exoproteins operated in eight different major biological pathways, namely putrescine biosynthesis, intracellular protein transport, membrane assembly, protein folding, metabolism, carbohydrate metabolism, proteolysis, and oxidation-reduction processes (Fig. 5).

As these results slightly differed from previous observations on the core exoproteome of a different and smaller set of samples (53), mainly for the lack of a pathogenesis GO term cluster, we also analyzed the

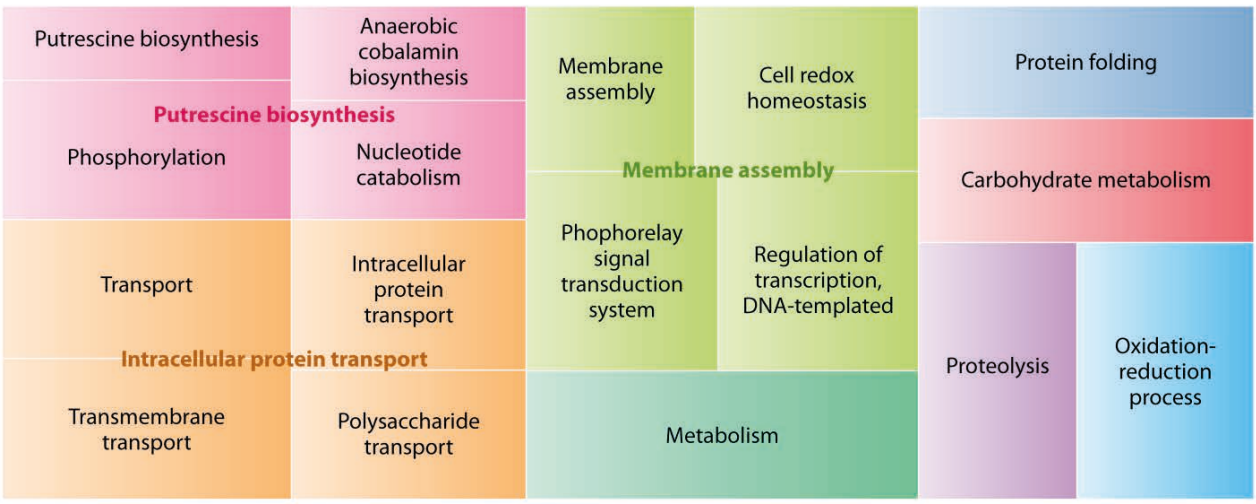


Figure 5. Biological pathways represented in the *P. gingivalis* core exoproteome. The REVIGO treemap depicts the outcome of a GO term analysis of cellular pathways involving the proteins predicted to define the core exoproteome of the *P. gingivalis* strains under examination.

variable exoproteome (Table S7). The simple absence of one virulence factor from one strain, in fact, would eliminate the protein from the core exoproteome and relegate it to the variable exoproteome. As expected, the GO term analyses of the variable exoproteomes revealed a sizable amount of extracellular proteins involved in pathogenesis in all *P. gingivalis* strains, except MDS140. The latter strain happens to be isolated from a healthy carrier. It is therefore tempting to speculate that the MDS140 strain could lack a number of virulence factors. Additionally, only the 'transport' and 'proteolysis' labels were assigned to predicted exoproteins of the MDS140 isolate (Fig. 6A), in contrast to the various other functional labels assigned to exoproteins from the other investigated strains (Fig. 6B-D).

Lastly, all the protein sorting information gathered by reviewing the available literature and predicting subcellular protein localization in *P. gingivalis* has been combined in Figure 7, which presents the total numbers of proteins predicted per subcellular compartment. Of note, this overview image distinguishes the core and variant proteomes of each compartment.

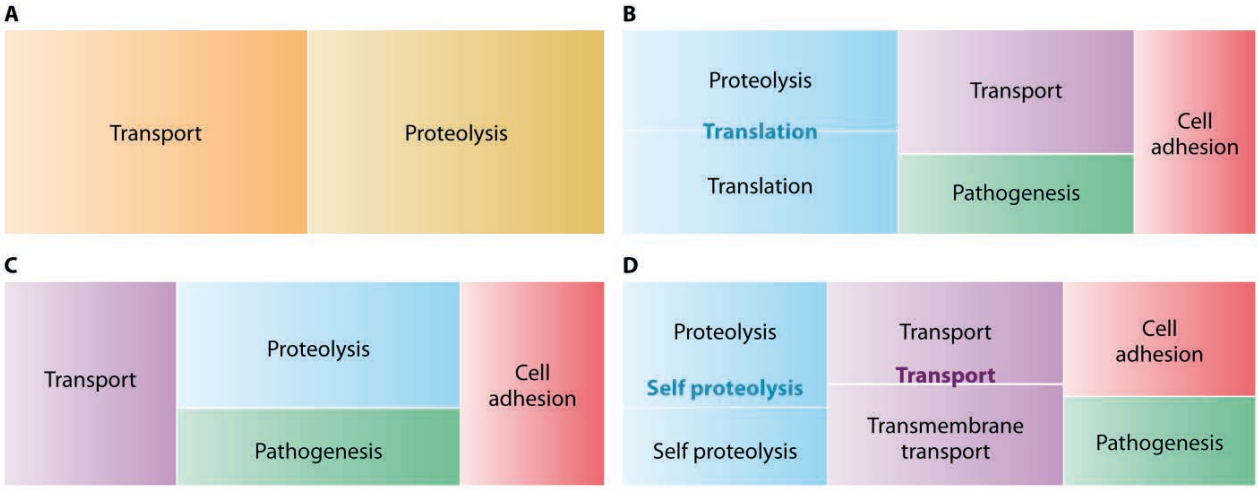


Figure 6. Biological pathways represented by the *P. gingivalis* variable exoproteome. The REVIGO treemaps represent the outcomes of GO term analyses of the cellular pathways in which the variable exoproteomes of different *P. gingivalis* strains are involved: **A)** MDS140; **B)** W83; **C)** TDC60, MDS33, 512915; **D)** ATCC 33277, 20655.

Conclusion

This review is focused on protein localization in the oral pathogen *P. gingivalis.* It integrates the results of published biochemical studies (51, 59-61) and a tailored *in silico* evaluation of published genome sequences that is grounded on established bioinformatic approaches (129, 130, 137). Considering the broad spectrum of interests that *P. gingivalis* elicits, especially in the fields of periodontology, rheumatology, and microbiology, this review will serve as an important lead for many upcoming studies concerning this bacterium. In fact, a compendium of the different subcellular and extracellular destination(s) that each individual protein may reach constitutes a treasure trove of invaluable information for any kind of research involving the biology and virulence of this bacterium. This view is underscored by the importance of the exoproteome in bacterial virulence, adhesion, and biofilm development, as well as diagnostic and therapeutic applications. The presently highlighted pathways for subcellular protein localization and secretion combined with the predicted protein addresses in *P. gingivalis* – in short, the 'Gingimaps' – could therefore be used to devise diagnostic or therapeutic antibodies targeting specific surface proteins, to create vaccines, and to discover druggable targets. This view is supported by the finding that, in a mouse model, oral infection by *P. gingivalis* and bacterial dissemination to arthritic joints can be inhibited with an anti-FimA antibody (165). Additionally,

as several proteins of *P. gingivalis* are subject of ongoing studies, the availability of data regarding the proteins belonging to the same subcellular compartments is a significant advantage when looking for targets, inhibitors, or possible cofactors. A simple example of this is utilizing exoproteomic data to narrow down the list of possible targets of the citrullinating enzyme PPAD. The same can be applied to gingipains, whose high proteolytic potential is under investigation in multiple fields, both clinical and biochemical.

Lastly, all the known and yet unknown mechanisms responsible for *P. gingivalis*' status as a successful oral pathogen implicated in a variety of diseases rely directly or indirectly on proteins. A direct impact of *P. gingivalis* proteins in disease is highlighted by the biological functions of PPAD, gingipains, hemagglutinins, and fimbriae, but there are likely to be many more. The indirect relationships of *P. gingivalis* proteins with human diseases are underpinned by the machinery needed to synthesize lipopolysaccharides and capsular components. Consequently, the present 'Gingimaps' may hold the key to a better understanding of causal or indirect relationships between this bacterium and the disorders to which is linked.

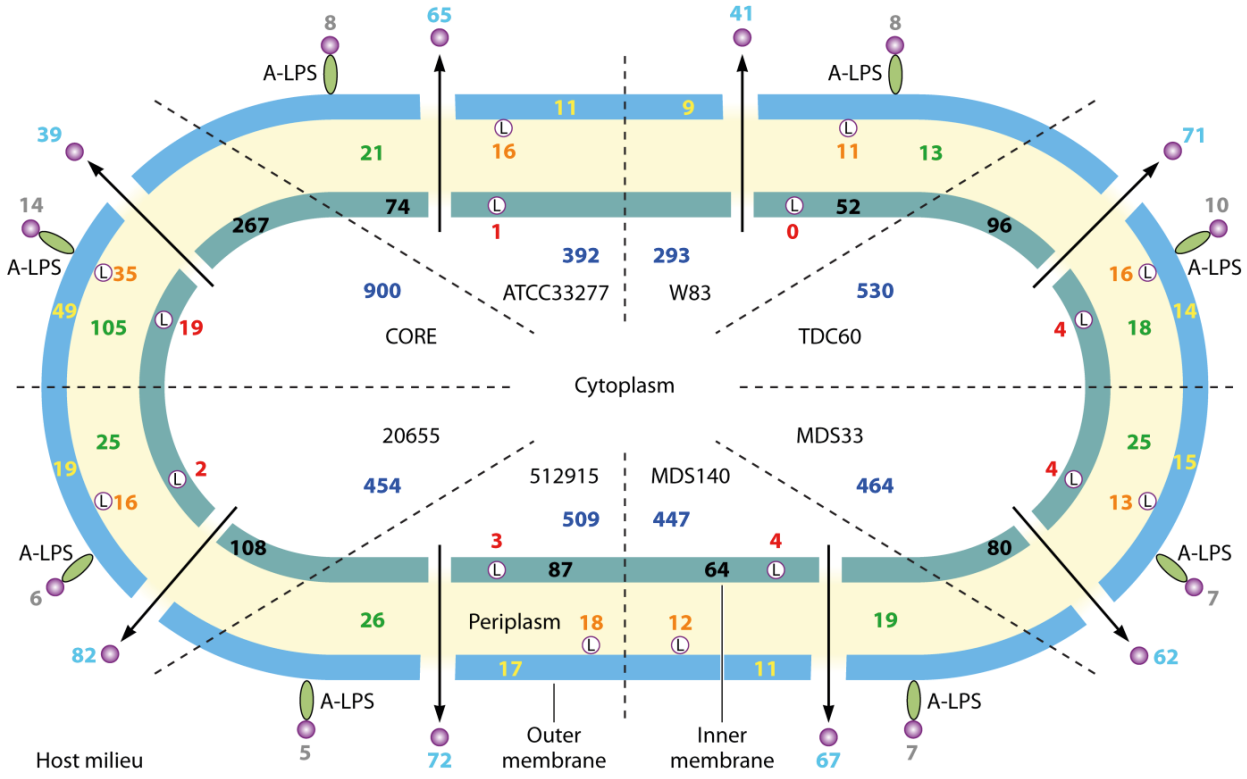


Figure 7. Overview of the subcellular localization of core and variant *P. gingivalis* proteins. The numbers of proteins residing at a particular subcellular location, or extracellularly, are indicated for the core proteome and the variable proteome of each examined *P. gingivalis* strain. Specifically, these include strains ATCC 33277, W83, TDC60, MDS33, MDS140, 512915 and 20655. The numbers of cytoplasmic proteins are indicated in blue, IM lipoproteins in red, IM proteins in black, periplasmic proteins in green, OM lipoproteins in orange, OM proteins in yellow, PorSS secreted proteins in grey and ECP-secreted proteins in cyan.

Acknowledgements

This work was supported by the Graduate School of Medical Sciences of the University of Groningen [to G.G.], the Center for Dentistry and Oral Hygiene of the University Medical Center Groningen [to G.G., A.J.v.W.], and the European Union's Horizon 2020 Programme under REA grant agreement no. 642836 [to S.G., J.M.v.D.]. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

The authors declare that they have no financial and non-financial competing interests in relation to the documented research.

References

1. Lauber, F., J. C. Deme, S. M. Lea, and B. C. Berks. 2018. Type 9 secretion system structures reveal a new protein transport mechanism. Nature. 564:77-82. doi: 10.1038/s41586-018-0693-y [doi].

2. Dominy, S. S., C. Lynch, F. Ermini, M. Benedyk, A. Marczyk, A. Konradi, M. Nguyen, U. Haditsch, D. Raha, C. Griffin, L. J. Holsinger, S. Arastu-Kapur, S. Kaba, A. Lee, M. I. Ryder, B. Potempa, P. Mydel, A. Hellvard, K. Adamowicz, H. Hasturk, G. D. Walker, E. C. Reynolds, R. L. M. Faull, M. A. Curtis, M. Dragunow, and J. Potempa. 2019. *Porphyromonas gingivalis* in Alzheimer's disease brains: Evidence for disease causation and treatment with small-molecule inhibitors. Science Advances. 5:eaau3333. doi: 10.1126/sciadv.aau3333.

3. van Winkelhoff, A. J., B. G. Loos, W. A. van der Reijden, and U. van der Velden. 2002. *Porphyromonas gingivalis, Bacteroides forsythus* and other putative periodontal pathogens in subjects with and without periodontal destruction. J. Clin. Periodontol. 29:1023-1028. doi: cpe291107 [pii].

4. Bostanci, N., and G. N. Belibasakis. 2012. *Porphyromonas gingivalis*: an invasive and evasive opportunistic oral pathogen. FEMS Microbiol. Lett. 333:1-9. doi: 10.1111/j.1574-6968.2012.02579.x [doi].

5. Yang, H. W., Y. F. Huang, and M. Y. Chou. 2004. Occurrence of *Porphyromonas gingivalis* and *Tannerella forsythensis* in periodontally diseased and healthy subjects. J. Periodontol. 75:1077-1083. doi: 10.1902/jop.2004.75.8.1077 [doi].

6. Datta, H. K., W. F. Ng, J. A. Walker, S. P. Tuck, and S. S. Varanasi. 2008. The cell biology of bone metabolism. J. Clin. Pathol. 61:577-587. doi: 10.1136/jcp.2007.048868 [doi].

7. How, K. Y., K. P. Song, and K. G. Chan. 2016. *Porphyromonas gingivalis*: An Overview of Periodontopathic Pathogen below the Gum Line. Front. Microbiol. 7:53. doi: 10.3389/fmicb.2016.00053 [doi].

8. Rylev, M., and M. Kilian. 2008. Prevalence and distribution of principal periodontal pathogens worldwide. J. Clin. Periodontol. 35:346-361. doi: 10.1111/j.1600-051X.2008.01280.x [doi].

9. Potempa, J., P. Mydel, and J. Koziel. 2017. The case for periodontitis in the pathogenesis of rheumatoid arthritis. Nat. Rev. Rheumatol. . doi: 10.1038/nrrheum.2017.132 [doi].

10. Darveau, R. P. 2010. Periodontitis: a polymicrobial disruption of host homeostasis. Nat. Rev. Microbiol. 8:481-490. doi: 10.1038/nrmicro2337 [doi].

11. Zhu, M., and B. S. Nikolajczyk. 2014. Immune cells link obesity-associated type 2 diabetes and periodontitis. J. Dent. Res. 93:346-352. doi: 10.1177/0022034513518943 [doi].

12. Chee, B., B. Park, and P. M. Bartold. 2013. Periodontitis and type II diabetes: a two-way relationship. Int. J. Evid Based. Healthc. 11:317-329. doi: 10.1111/1744-1609.12038 [doi].

13. Preshaw, P. M., and S. M. Bissett. 2013. Periodontitis: oral complication of diabetes. Endocrinol. Metab. Clin. North Am. 42:849-867. doi: 10.1016/j.ecl.2013.05.012 [doi].

14. Kumar, M., L. Mishra, R. Mohanty, and R. Nayak. 2014. "Diabetes and gum disease: the diabolic duo". Diabetes Metab. Syndr. 8:255-258. doi: 10.1016/j.dsx.2014.09.022 [doi].

15. Bascones-Martinez, A., J. Gonzalez-Febles, and J. Sanz-Esporrin. 2014. Diabetes and periodontal disease. Review of the literature. Am. J. Dent. 27:63-67.

16. Gurav, A. N. 2014. The association of periodontitis and metabolic syndrome. Dent. Res. J. (Isfahan). 11:1-10.

17. Kebschull, M., R. T. Demmer, and P. N. Papapanou. 2010. "Gum bug, leave my heart alone!"--epidemiologic and mechanistic evidence linking periodontal infections and atherosclerosis. J. Dent. Res. 89:879-902. doi: 10.1177/0022034510375281 [doi].

18. Kelly, J. T., G. Avila-Ortiz, V. Allareddy, G. K. Johnson, and S. Elangovan. 2013. The association between periodontitis and coronary heart disease: a quality assessment of systematic reviews. J. Am. Dent. Assoc. 144:371-379. doi: 144/4/371 [pii].

19. Li, C., Z. Lv, Z. Shi, Y. Zhu, Y. Wu, L. Li, and Z. Iheozor-Ejiofor. 2014. Periodontal therapy for the management

of cardiovascular disease in patients with chronic periodontitis. Cochrane Database Syst. Rev. 8:CD009197. doi: 10.1002/14651858.CD009197.pub2 [doi].
20. Noble, J. M., N. Scarmeas, and P. N. Papapanou. 2013. Poor oral health as a chronic, potentially modifiable dementia risk factor: review of the literature. Curr. Neurol. Neurosci. Rep. 13:384-013-0384-x. doi: 10.1007/s11910-013-0384-x [doi].
21. Shaik, M. M., S. Ahmad, S. H. Gan, A. M. Abuzenadah, E. Ahmad, S. Tabrez, F. Ahmed, and M. A. Kamal. 2014. How do periodontal infections affect the onset and progression of Alzheimer's disease? CNS Neurol. Disord. Drug Targets. 13:460-466. doi: CNSNDDT-EPUB-56157 [pii].
22. Kamer, A. R., R. G. Craig, A. P. Dasanayake, M. Brys, L. Glodzik-Sobanska, and M. J. de Leon. 2008. Inflammation and Alzheimer's disease: possible role of periodontal diseases. Alzheimers Dement. 4:242-250. doi: 10.1016/j.jalz.2007.08.004 [doi].
23. Kamer, A. R., A. P. Dasanayake, R. G. Craig, L. Glodzik-Sobanska, M. Bry, and M. J. de Leon. 2008. Alzheimer's disease and peripheral infections: the possible contribution from periodontal infections, model and hypothesis. J. Alzheimers Dis. 13:437-449.
24. de Pablo, P., T. Dietrich, and T. E. McAlindon. 2008. Association of periodontal disease and tooth loss with rheumatoid arthritis in the US population. J. Rheumatol. 35:70-76. doi: 07/13/1123 [pii].
25. de Smit, M., J. Westra, A. Vissink, B. Doornbos-van der Meer, E. Brouwer, and A. J. van Winkelhoff. 2012. Periodontitis in established rheumatoid arthritis patients: a cross-sectional clinical, microbiological and serological study. Arthritis Res. Ther. 14:R222. doi: 10.1186/ar4061 [doi].
26. de Smit, M. J., E. Brouwer, J. Westra, W. Nesse, A. Vissink, and A. J. van Winkelhoff. 2012. Effect of periodontal treatment on rheumatoid arthritis and vice versa. Ned. Tijdschr. Tandheelkd. 119:191-197.
27. de Smit, M. J., E. Brouwer, A. Vissink, and A. J. van Winkelhoff. 2011. Rheumatoid arthritis and periodontitis; a possible link via citrullination. Anaerobe. 17:196-200. doi: 10.1016/j.anaerobe.2011.03.019 [doi].
28. Detert, J., N. Pischon, G. R. Burmester, and F. Buttgereit. 2010. The association between rheumatoid arthritis and periodontal disease. Arthritis Res. Ther. 12:218. doi: 10.1186/ar3106 [doi].
29. Dissick, A., R. S. Redman, M. Jones, B. V. Rangan, A. Reimold, G. R. Griffiths, T. R. Mikuls, R. L. Amdur, J. S. Richards, and G. S. Kerr. 2010. Association of periodontitis with rheumatoid arthritis: a pilot study. J. Periodontol. 81:223-230. doi: 10.1902/jop.2009.090309 [doi].
30. El-Shinnawi, U., and M. Soory. 2013. Associations between periodontitis and systemic inflammatory diseases: response to treatment. Recent. Pat. Endocr Metab. Immune Drug Discov. 7:169-188. doi: 54959 [pii].
31. Farquharson, D., J. P. Butcher, and S. Culshaw. 2012. Periodontitis, *Porphyromonas*, and the pathogenesis of rheumatoid arthritis. Mucosal Immunol. 5:112-120. doi: 10.1038/mi.2011.66 [doi].
32. Hitchon, C. A., and H. S. El-Gabalawy. 2011. Infection and rheumatoid arthritis: still an open question. Curr. Opin. Rheumatol. 23:352-357. doi: 10.1097/BOR.0b013e3283477b7b [doi].
33. Lundberg, K., N. Wegner, T. Yucel-Lindberg, and P. J. Venables. 2010. Periodontitis in RA-the citrullinated enolase connection. Nat. Rev. Rheumatol. 6:727-730. doi: 10.1038/nrrheum.2010.139 [doi].
34. Maresz, K. J., A. Hellvard, A. Sroka, K. Adamowicz, E. Bielecka, J. Koziel, K. Gawron, D. Mizgalska, K. A. Marcinska, M. Benedyk, K. Pyrc, A. M. Quirke, R. Jonsson, S. Alzabin, P. J. Venables, K. A. Nguyen, P. Mydel, and J. Potempa. 2013. *Porphyromonas gingivalis* facilitates the development and progression of destructive arthritis through its unique bacterial peptidylarginine deiminase (PAD). PLoS Pathog. 9:e1003627. doi: 10.1371/journal.ppat.1003627 [doi].
35. Mercado, F., R. I. Marshall, A. C. Klestov, and P. M. Bartold. 2000. Is there a relationship between rheumatoid arthritis and periodontal disease? J. Clin. Periodontol. 27:267-272.
36. Mercado, F. B., R. I. Marshall, A. C. Klestov, and P. M. Bartold. 2001. Relationship between rheumatoid arthritis and periodontitis. J. Periodontol. 72:779-787. doi: 10.1902/jop.2001.72.6.779 [doi].
37. Pischon, N., T. Pischon, J. Kroger, E. Gulmez, B. M. Kleber, J. P. Bernimoulin, H. Landau, P. G. Brinkmann,

P. Schlattmann, J. Zernicke, F. Buttgereit, and J. Detert. 2008. Association among rheumatoid arthritis, oral hygiene, and periodontitis. J. Periodontol. 79:979-986. doi: 10.1902/jop.2008.070501 [doi].

38. Quirke, A. M., E. B. Lugli, N. Wegner, B. C. Hamilton, P. Charles, M. Chowdhury, A. J. Ytterberg, R. A. Zubarev, J. Potempa, S. Culshaw, Y. Guo, B. A. Fisher, G. Thiele, T. R. Mikuls, and P. J. Venables. 2014. Heightened immune response to autocitrullinated *Porphyromonas gingivalis* peptidylarginine deiminase: a potential mechanism for breaching immunologic tolerance in rheumatoid arthritis. Ann. Rheum. Dis. 73:263-269. doi: 10.1136/annrheumdis-2012-202726 [doi].

39. Routsias, J. G., J. D. Goules, A. Goules, G. Charalampakis, and D. Pikazis. 2011. Autopathogenic correlation of periodontitis and rheumatoid arthritis. Rheumatology (Oxford). 50:1189-1193. doi: 10.1093/rheumatology/ker090 [doi].

40. Tolo, K., and L. Jorkjend. 1990. Serum antibodies and loss of periodontal bone in patients with rheumatoid arthritis. J. Clin. Periodontol. 17:288-291.

41. Wegner, N., K. Lundberg, A. Kinloch, B. Fisher, V. Malmstrom, M. Feldmann, and P. J. Venables. 2010. Autoimmunity to specific citrullinated proteins gives the first clues to the etiology of rheumatoid arthritis. Immunol. Rev. 233:34-54. doi: 10.1111/j.0105-2896.2009.00850.x [doi].

42. Wegner, N., R. Wait, A. Sroka, S. Eick, K. A. Nguyen, K. Lundberg, A. Kinloch, S. Culshaw, J. Potempa, and P. J. Venables. 2010. Peptidylarginine deiminase from *Porphyromonas gingivalis* citrullinates human fibrinogen and alpha-enolase: implications for autoimmunity in rheumatoid arthritis. Arthritis Rheum. 62:2662-2672. doi: 10.1002/art.27552 [doi].

43. Kasser, U. R., C. Gleissner, F. Dehne, A. Michel, B. Willershausen-Zonnchen, and W. W. Bolten. 1997. Risk for periodontal disease in patients with longstanding rheumatoid arthritis. Arthritis Rheum. 40:2248-2251. doi: 10.1002/1529-0131(199712)40:12<2248::AID-ART20>3.0.CO;2-W [doi].

44. Nesse, W., P. U. Dijkstra, F. Abbas, F. K. Spijkervet, A. Stijger, J. A. Tromp, J. L. van Dijk, and A. Vissink. 2010. Increased prevalence of cardiovascular and autoimmune diseases in periodontitis patients: a cross-sectional study. J. Periodontol. 81:1622-1628. doi: 10.1902/jop.2010.100058 [doi].

45. Mangat, P., N. Wegner, P. J. Venables, and J. Potempa. 2010. Bacterial and human peptidylarginine deiminases: targets for inhibiting the autoimmune response in rheumatoid arthritis? Arthritis Res. Ther. 12:209. doi: 10.1186/ar3000 [doi].

46. Gabarrini, G., M. de Smit, J. Westra, E. Brouwer, A. Vissink, K. Zhou, J. W. Rossen, T. Stobernack, J. M. van Dijl, and A. J. van Winkelhoff. 2015. The peptidylarginine deiminase gene is a conserved feature of *Porphyromonas gingivalis*. Sci. Rep. 5:13936. doi: 10.1038/srep13936 [doi].

47. Gabarrini, G., M. A. Chlebowicz, M. E. Vega Quiroz, A. C. M. Veloo, J. W. A. Rossen, H. J. M. Harmsen, M. L. Laine, J. M. van Dijl, and A. J. van Winkelhoff. 2017. Conserved Citrullinating Exoenzymes in *Porphyromonas* Species. J. Dent. Res. 22034517747575. doi: 10.1177/0022034517747575 [doi].

48. Goulas, T., D. Mizgalska, I. Garcia-Ferrer, T. Kantyka, T. Guevara, B. Szmigielski, A. Sroka, C. Millan, I. Uson, F. Veillard, B. Potempa, P. Mydel, M. Sola, J. Potempa, and F. X. Gomis-Ruth. 2015. Structure and mechanism of a bacterial host-protein citrullinating virulence factor, *Porphyromonas gingivalis* peptidylarginine deiminase. Sci. Rep. 5:11969. doi: 10.1038/srep11969 [doi].

49. Avouac, J., L. Gossec, and M. Dougados. 2006. Diagnostic and predictive value of anti-cyclic citrullinated protein antibodies in rheumatoid arthritis: a systematic literature review. Ann. Rheum. Dis. 65:845-851. doi: ard.2006.051391 [pii].

50. Nishimura, K., D. Sugiyama, Y. Kogata, G. Tsuji, T. Nakazawa, S. Kawano, K. Saigo, A. Morinobu, M. Koshiba, K. M. Kuntz, I. Kamae, and S. Kumagai. 2007. Meta-analysis: diagnostic accuracy of anti-cyclic citrullinated peptide antibody and rheumatoid factor for rheumatoid arthritis. Ann. Intern. Med. 146:797-808. doi: 146/11/797 [pii].

51. Sato, K., H. Yukitake, Y. Narita, M. Shoji, M. Naito, and K. Nakayama. 2013. Identification of *Porphyromonas*

gingivalis proteins secreted by the Por secretion system. FEMS Microbiol. Lett. 338:68-76. doi: 10.1111/1574-6968.12028 [doi].

52. Gabarrini, G., L. M. Palma Medina, T. Stobernack, R. C. Prins, M. du Teil Espina, J. Kuipers, M. A. Chlebowicz, J. W. Rossen, A. J. van Winkelhoff, and J. M. van Dijl. 2018. There's no place like OM: Vesicular sorting and secretion of the peptidylarginine deiminase of *Porphyromonas gingivalis*. Virulence. 9:456-464.

53. Stobernack, T., C. Glasner, S. Junker, G. Gabarrini, M. de Smit, A. de Jong, A. Otto, D. Becher, A. J. van Winkelhoff, and J. M. van Dijl. 2016. Extracellular Proteome and Citrullinome of the Oral Pathogen *Porphyromonas gingivalis*. J. Proteome Res. 15:4532-4543. doi: 10.1021/acs.jproteome.6b00634 [doi].

54. Laine, M. L., B. J. Appelmelk, and A. J. van Winkelhoff. 1997. Prevalence and distribution of six capsular serotypes of *Porphyromonas gingivalis* in periodontitis patients. J. Dent. Res. 76:1840-1844.

55. Brunner, J., N. Scheres, N. B. El Idrissi, D. M. Deng, M. L. Laine, A. J. van Winkelhoff, and W. Crielaard. 2010. The capsule of *Porphyromonas gingivalis* reduces the immune response of human gingival fibroblasts. BMC Microbiol. 10:5-2180-10-5. doi: 10.1186/1471-2180-10-5 [doi].

56. Lamont, R. J., and H. F. Jenkinson. 1998. Life below the gum line: pathogenic mechanisms of *Porphyromonas gingivalis*. Microbiol. Mol. Biol. Rev. 62:1244-1263.

57. Honda, K. 2011. *Porphyromonas gingivalis* sinks teeth into the oral microbiota and periodontal disease. Cell. Host Microbe. 10:423-425. doi: 10.1016/j.chom.2011.10.008 [doi].

58. Hajishengallis, G., R. P. Darveau, and M. A. Curtis. 2012. The keystone-pathogen hypothesis. Nat. Rev. Microbiol. 10:717-725. doi: 10.1038/nrmicro2873 [doi].

59. Veith, P. D., G. H. Talbo, N. Slakeski, S. G. Dashper, C. Moore, R. A. Paolini, and E. C. Reynolds. 2002. Major outer membrane proteins and proteolytic processing of RgpA and Kgp of *Porphyromonas gingivalis* W50. Biochem. J. 363:105-115.

60. Yoshimura, F., Y. Murakami, K. Nishikawa, Y. Hasegawa, and S. Kawaminami. 2009. Surface components of *Porphyromonas gingivalis*. J. Periodontal. Res. 44:1-12. doi: 10.1111/j.1600-0765.2008.01135.x [doi].

61. McBride, M. J., and Y. Zhu. 2013. Gliding motility and Por secretion system genes are widespread among members of the phylum Bacteroidetes. J. Bacteriol. 195:270-278. doi: 10.1128/JB.01962-12 [doi].

62. Scheres, N., R. J. Lamont, W. Crielaard, and B. P. Krom. 2015. LuxS signaling in *Porphyromonas gingivalis*-host interactions. Anaerobe. 35:3-9. doi: 10.1016/j.anaerobe.2014.11.011 [doi].

63. Glew, M. D., P. D. Veith, D. Chen, C. A. Seers, Y. Y. Chen, and E. C. Reynolds. 2014. Blue native-PAGE analysis of membrane protein complexes in *Porphyromonas gingivalis*. J. Proteomics. 110:72-92. doi: 10.1016/j.jprot.2014.07.033 [doi].

64. Rangarajan, M., J. Aduse-Opoku, A. Hashim, G. McPhail, Z. Luklinska, M. F. Haurat, M. F. Feldman, and M. A. Curtis. 2017. LptO (PG0027) Is Required for Lipid A 1-Phosphatase Activity in *Porphyromonas gingivalis* W50. J. Bacteriol. 199:10.1128/JB.00751-16. Print 2017 Jun 1. doi: e00751-16 [pii].

65. Nonaka, M., M. Shoji, T. Kadowaki, K. Sato, H. Yukitake, M. Naito, and K. Nakayama. 2014. Analysis of a Lys-specific serine endopeptidase secreted via the type IX secretion system in *Porphyromonas gingivalis*. FEMS Microbiol. Lett. 354:60-68. doi: 10.1111/1574-6968.12426 [doi].

66. Saiki, K., and K. Konishi. 2014. *Porphyromonas gingivalis* C-terminal signal peptidase PG0026 and HagA interact with outer membrane protein PG27/LptO. Mol. Oral Microbiol. 29:32-44. doi: 10.1111/omi.12043 [doi].

67. Hasegawa, Y., K. Nagano, R. Ikai, M. Izumigawa, Y. Yoshida, N. Kitai, R. J. Lamont, Y. Murakami, and F. Yoshimura. 2013. Localization and function of the accessory protein Mfa3 in *Porphyromonas gingivalis* Mfa1 fimbriae. Mol. Oral Microbiol. 28:467-480. doi: 10.1111/omi.12040 [doi].

68. Mantri, C. K., C. H. Chen, X. Dong, J. S. Goodwin, S. Pratap, V. Paromov, and H. Xie. 2015. Fimbriae-mediated outer membrane vesicle production and invasion of *Porphyromonas gingivalis*. Microbiologyopen. 4:53-65. doi: 10.1002/mbo3.221 [doi].

69. Nagano, K., Y. Hasegawa, Y. Yoshida, and F. Yoshimura. 2015. A Major Fimbrilin Variant of Mfa1 Fimbriae in *Porphyromonas gingivalis*. J. Dent. Res. 94:1143-1148. doi: 10.1177/0022034515588275 [doi].
70. Staniec, D., M. Ksiazek, I. B. Thogersen, J. J. Enghild, A. Sroka, D. Bryzek, M. Bogyo, M. Abrahamson, and J. Potempa. 2015. Calcium Regulates the Activity and Structural Stability of Tpr, a Bacterial Calpain-like Peptidase. J. Biol. Chem. 290:27248-27260. doi: 10.1074/jbc.M115.648782 [doi].
71. Gorasia, D. G., P. D. Veith, D. Chen, C. A. Seers, H. A. Mitchell, Y. Y. Chen, M. D. Glew, S. G. Dashper, and E. C. Reynolds. 2015. *Porphyromonas gingivalis* Type IX Secretion Substrates Are Cleaved and Modified by a Sortase-Like Mechanism. PLoS Pathog. 11:e1005152. doi: 10.1371/journal.ppat.1005152 [doi].
72. Ota, K., Y. Kikuchi, K. Imamura, D. Kita, K. Yoshikawa, A. Saito, and K. Ishihara. 2017. SigCH, an extracytoplasmic function sigma factor of *Porphyromonas gingivalis* regulates the expression of cdhR and hmuYR. Anaerobe. 43:82-90. doi: S1075-9964(16)30165-2 [pii].
73. Lasica, A. M., T. Goulas, D. Mizgalska, X. Zhou, I. de Diego, M. Ksiazek, M. Madej, Y. Guo, T. Guevara, M. Nowak, B. Potempa, A. Goel, M. Sztukowska, A. T. Prabhakar, M. Bzowska, M. Widziolek, I. B. Thogersen, J. J. Enghild, M. Simonian, A. W. Kulczyk, K. A. Nguyen, J. Potempa, and F. X. Gomis-Ruth. 2016. Structural and functional probing of PorZ, an essential bacterial surface component of the type-IX secretion system of human oral-microbiomic *Porphyromonas gingivalis*. Sci. Rep. 6:37708. doi: 10.1038/srep37708 [doi].
74. Heath, J. E., C. A. Seers, P. D. Veith, C. A. Butler, N. A. Nor Muhammad, Y. Y. Chen, N. Slakeski, B. Peng, L. Zhang, S. G. Dashper, K. J. Cross, S. M. Cleal, C. Moore, and E. C. Reynolds. 2016. PG1058 Is a Novel Multidomain Protein Component of the Bacterial Type IX Secretion System. PLoS One. 11:e0164313. doi: 10.1371/journal.pone.0164313 [doi].
75. Goulas, T., I. Garcia-Ferrer, J. A. Hutcherson, B. A. Potempa, J. Potempa, D. A. Scott, and F. Xavier Gomis-Ruth. 2016. Structure of RagB, a major immunodominant outer-membrane surface receptor antigen of *Porphyromonas gingivalis*. Mol. Oral Microbiol. 31:472-485. doi: 10.1111/omi.12140 [doi].
76. Naylor, K. L., M. Widziolek, S. Hunt, M. Conolly, M. Hicks, P. Stafford, J. Potempa, C. Murdoch, C. W. Douglas, and G. P. Stafford. 2017. Role of OmpA2 surface regions of *Porphyromonas gingivalis* in host-pathogen interactions with oral epithelial cells. Microbiologyopen. 6:10.1002/mbo3.401. Epub 2016 Sep 6. doi: 10.1002/mbo3.401 [doi].
77. Gorasia, D. G., P. D. Veith, E. G. Hanssen, M. D. Glew, K. Sato, H. Yukitake, K. Nakayama, and E. C. Reynolds. 2016. Structural Insights into the PorK and PorN Components of the *Porphyromonas gingivalis* Type IX Secretion System. PLoS Pathog. 12:e1005820. doi: 10.1371/journal.ppat.1005820 [doi].
78. Hasegawa, Y., Y. Iijima, K. Persson, K. Nagano, Y. Yoshida, R. J. Lamont, T. Kikuchi, A. Mitani, and F. Yoshimura. 2016. Role of Mfa5 in Expression of Mfa1 Fimbriae in *Porphyromonas gingivalis*. J. Dent. Res. 95:1291-1297. doi: 10.1177/0022034516655083 [doi].
79. Smalley, J. W., and T. Olczak. 2017. Heme acquisition mechanisms of *Porphyromonas gingivalis* - strategies used in a polymicrobial community in a heme-limited host environment. Mol. Oral Microbiol. 32:1-23. doi: 10.1111/omi.12149 [doi].
80. Taguchi, Y., K. Sato, H. Yukitake, T. Inoue, M. Nakayama, M. Naito, Y. Kondo, K. Kano, T. Hoshino, K. Nakayama, S. Takashiba, and N. Ohara. 2015. Involvement of an Skp-Like Protein, PGN_0300, in the Type IX Secretion System of *Porphyromonas gingivalis*. Infect. Immun. 84:230-240. doi: 10.1128/IAI.01308-15 [doi].
81. Slakeski, N., S. G. Dashper, P. Cook, C. Poon, C. Moore, and E. C. Reynolds. 2000. A *Porphyromonas gingivalis* genetic locus encoding a heme transport system. Oral Microbiol. Immunol. 15:388-392. doi: omi150609 [pii].
82. Karunakaran, T., T. Madden, and H. Kuramitsu. 1997. Isolation and characterization of a hemin-regulated gene, hemR, from *Porphyromonas gingivalis*. J. Bacteriol. 179:1898-1908.
83. Fujise, K., Y. Kikuchi, E. Kokubu, K. Okamoto-Shibayama, and K. Ishihara. 2017. Effect of extracytoplasmic function sigma factors on autoaggregation, hemagglutination, and cell surface properties of *Porphyromonas*

gingivalis. PLoS One. 12:e0185027. doi: 10.1371/journal.pone.0185027 [doi].
84. Veith, P. D., M. D. Glew, D. G. Gorasia, and E. C. Reynolds. 2017. Type IX secretion: the generation of bacterial cell surface coatings involved in virulence, gliding motility and the degradation of complex biopolymers. Mol. Microbiol. 106:35-53. doi: 10.1111/mmi.13752 [doi].
85. Lasica, A. M., M. Ksiazek, M. Madej, and J. Potempa. 2017. The Type IX Secretion System (T9SS): Highlights and Recent Insights into Its Structure and Function. Front. Cell. Infect. Microbiol. 7:215. doi: 10.3389/fcimb.2017.00215 [doi].
86. The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45:D158-D169. doi: 10.1093/nar/gkw1099 [doi].
87. Hoover, C. I., E. Abarbarchuk, C. Y. Ng, and J. R. Felton. 1992. Transposition of Tn4351 in *Porphyromonas gingivalis*. Plasmid. 27:246-250.
88. Sandmeier, H., K. Bar, and J. Meyer. 1993. Search for bacteriophages of black-pigmented gram-negative anaerobes from dental plaque. FEMS Immunol. Med. Microbiol. 6:193-194.
89. Veith, P. D., Y. Y. Chen, D. G. Gorasia, D. Chen, M. D. Glew, N. M. O'Brien-Simpson, J. D. Cecil, J. A. Holden, and E. C. Reynolds. 2014. *Porphyromonas gingivalis* outer membrane vesicles exclusively contain outer membrane and periplasmic proteins and carry a cargo enriched with virulence factors. J. Proteome Res. 13:2420-2432. doi: 10.1021/pr401227e [doi].
90. Gui, M. J., S. G. Dashper, N. Slakeski, Y. Y. Chen, and E. C. Reynolds. 2016. Spheres of influence: *Porphyromonas gingivalis* outer membrane vesicles. Mol. Oral Microbiol. 31:365-378. doi: 10.1111/omi.12134 [doi].
91. Ellis, T. N., and M. J. Kuehn. 2010. Virulence and immunomodulatory roles of bacterial outer membrane vesicles. Microbiol. Mol. Biol. Rev. 74:81-94. doi: 10.1128/MMBR.00031-09 [doi].
92. O'Brien-Simpson, N. M., R. D. Pathirana, G. D. Walker, and E. C. Reynolds. 2009. *Porphyromonas gingivalis* RgpA-Kgp proteinase-adhesin complexes penetrate gingival tissue and induce proinflammatory cytokines or apoptosis in a concentration-dependent manner. Infect. Immun. 77:1246-1261. doi: 10.1128/IAI.01038-08 [doi].
93. Xie, H. 2015. Biogenesis and function of *Porphyromonas gingivalis* outer membrane vesicles. Future Microbiol. 10:1517-1527. doi: 10.2217/fmb.15.63 [doi].
94. Furuta, N., H. Takeuchi, and A. Amano. 2009. Entry of *Porphyromonas gingivalis* outer membrane vesicles into epithelial cells causes cellular functional impairment. Infect. Immun. 77:4761-4770. doi: 10.1128/IAI.00841-09 [doi].
95. Furuta, N., K. Tsuda, H. Omori, T. Yoshimori, F. Yoshimura, and A. Amano. 2009. *Porphyromonas gingivalis* outer membrane vesicles enter human epithelial cells via an endocytic pathway and are sorted to lysosomal compartments. Infect. Immun. 77:4187-4196. doi: 10.1128/IAI.00009-09 [doi].
96. Zhu, Y., S. G. Dashper, Y. Y. Chen, S. Crawford, N. Slakeski, and E. C. Reynolds. 2013. *Porphyromonas gingivalis* and *Treponema denticola* synergistic polymicrobial biofilm development. PLoS One. 8:e71727. doi: 10.1371/journal.pone.0071727 [doi].
97. O-Brien-Simpson, N. M., P. D. Veith, S. G. Dashper, and E. C. Reynolds. 2003. *Porphyromonas gingivalis* gingipains: the molecular teeth of a microbial vampire. Curr. Protein Pept. Sci. 4:409-426.
98. Potempa, J., A. Banbula, and J. Travis. 2000. Role of bacterial proteinases in matrix destruction and modulation of host responses. Periodontol. 2000. 24:153-192.
99. Alvarez-Martinez, C. E., and P. J. Christie. 2009. Biological diversity of prokaryotic type IV secretion systems. Microbiol. Mol. Biol. Rev. 73:775-808. doi: 10.1128/MMBR.00023-09 [doi].
100. McBride, M. J., and D. Nakane. 2015. *Flavobacterium* gliding motility and the type IX secretion system. Curr. Opin. Microbiol. 28:72-77. doi: 10.1016/j.mib.2015.07.016 [doi].
101. Driessen, A. J., and N. Nouwen. 2008. Protein translocation across the bacterial cytoplasmic membrane. Annu. Rev. Biochem. 77:643-667. doi: 10.1146/annurev.biochem.77.061606.160747 [doi].

102. Sargent, F. 2007. The twin-arginine transport system: moving folded proteins across membranes. Biochem. Soc. Trans. 35:835-847. doi: BST0350835 [pii].
103. Robinson, C., C. F. Matos, D. Beck, C. Ren, J. Lawrence, N. Vasisht, and S. Mendel. 2011. Transport and proofreading of proteins by the twin-arginine translocation (Tat) system in bacteria. Biochim. Biophys. Acta. 1808:876-884. doi: 10.1016/j.bbamem.2010.11.023 [doi].
104. Goosens, V. J., and J. M. van Dijl. 2017. Twin-Arginine Protein Translocation. Curr. Top. Microbiol. Immunol. 404:69-94. doi: 10.1007/82_2016_7 [doi].
105. Knowles, T. J., A. Scott-Tucker, M. Overduin, and I. R. Henderson. 2009. Membrane protein architects: the role of the BAM complex in outer membrane protein assembly. Nat. Rev. Microbiol. 7:206-214. doi: 10.1038/nrmicro2069 [doi].
106. Konovalova, A., D. E. Kahne, and T. J. Silhavy. 2017. Outer Membrane Biogenesis. Annu. Rev. Microbiol. 71:539-556. doi: 10.1146/annurev-micro-090816-093754 [doi].
107. Narita, S., and H. Tokuda. 2006. An ABC transporter mediating the membrane detachment of bacterial lipoproteins depending on their sorting signals. FEBS Lett. 580:1164-1170. doi: S0014-5793(05)01298-6 [pii].
108. Okuda, S., and H. Tokuda. 2011. Lipoprotein sorting in bacteria. Annu. Rev. Microbiol. 65:239-259. doi: 10.1146/annurev-micro-090110-102859 [doi].
109. Rossiter, A. E., D. L. Leyton, K. Tveen-Jensen, D. F. Browning, Y. Sevastsyanovich, T. J. Knowles, K. B. Nichols, A. F. Cunningham, M. Overduin, M. A. Schembri, and I. R. Henderson. 2011. The essential beta-barrel assembly machinery complex components BamD and BamA are required for autotransporter biogenesis. J. Bacteriol. 193:4250-4253. doi: 10.1128/JB.00192-11 [doi].
110. Desvaux, M., M. Hebraud, R. Talon, and I. R. Henderson. 2009. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. Trends Microbiol. 17:139-145. doi: 10.1016/j.tim.2009.01.004 [doi].
111. Low, H. H., F. Gubellini, A. Rivera-Calzada, N. Braun, S. Connery, A. Dujeancourt, F. Lu, A. Redzej, R. Fronzes, E. V. Orlova, and G. Waksman. 2014. Structure of a type IV secretion system. Nature. 508:550-553. doi: 10.1038/nature13081 [doi].
112. Leone, P., J. Roche, M. S. Vincent, Q. H. Tran, A. Desmyter, E. Cascales, C. Kellenberger, C. Cambillau, and A. Roussel. 2018. Type IX secretion system PorM and gliding machinery GldM form arches spanning the periplasmic space. Nat. Commun. 9:429-017-02784-7. doi: 10.1038/s41467-017-02784-7 [doi].
113. Shoji, M., K. Sato, H. Yukitake, Y. Kondo, Y. Narita, T. Kadowaki, M. Naito, and K. Nakayama. 2011. Por secretion system-dependent secretion and glycosylation of *Porphyromonas gingivalis* hemin-binding protein 35. PLoS One. 6:e21372. doi: 10.1371/journal.pone.0021372 [doi].
114. Konig, M. F., A. S. Paracha, M. Moni, C. O. Bingham 3rd, and F. Andrade. 2015. Defining the role of *Porphyromonas gingivalis* peptidylarginine deiminase (PPAD) in rheumatoid arthritis through the study of PPAD biology. Ann. Rheum. Dis. 74:2054-2061. doi: annrheumdis-2014-205385 [pii].
115. Eichinger, A., H. G. Beisel, U. Jacob, R. Huber, F. J. Medrano, A. Banbula, J. Potempa, J. Travis, and W. Bode. 1999. Crystal structure of gingipain R: an Arg-specific bacterial cysteine proteinase with a caspase-like fold. EMBO J. 18:5453-5462. doi: 10.1093/emboj/18.20.5453 [doi].
116. Chen, Y. Y., K. J. Cross, R. A. Paolini, J. E. Fielding, N. Slakeski, and E. C. Reynolds. 2002. CPG70 is a novel basic metallocarboxypeptidase with C-terminal polycystic kidney disease domains from *Porphyromonas gingivalis*. J. Biol. Chem. 277:23433-23440. doi: 10.1074/jbc.M200811200 [doi].
117. Mikolajczyk, J., K. M. Boatright, H. R. Stennicke, T. Nazif, J. Potempa, M. Bogyo, and G. S. Salvesen. 2003. Sequential autolytic processing activates the zymogen of Arg-gingipain. J. Biol. Chem. 278:10458-10464. doi: 10.1074/jbc.M210564200 [doi].
118. Ngo, L. H., P. D. Veith, Y. Y. Chen, D. Chen, I. B. Darby, and E. C. Reynolds. 2010. Mass spectrometric analyses of peptides and proteins in human gingival crevicular fluid. J. Proteome Res. 9:1683-1693. doi: 10.1021/

pr900775s [doi].
119. Paramonov, N., D. Bailey, M. Rangarajan, A. Hashim, G. Kelly, M. A. Curtis, and E. F. Hounsell. 2001. Structural analysis of the polysaccharide from the lipopolysaccharide of *Porphyromonas gingivalis* strain W50. Eur. J. Biochem. 268:4698-4707. doi: ejb2397 [pii].
120. Stobernack, T., M. du Teil Espina, L. M. Mulder, L. M. Palma Medina, D. R. Piebenga, G. Gabarrini, X. Zhao, K. M. J. Janssen, J. Hulzebos, E. Brouwer, T. Sura, D. Becher, A. J. van Winkelhoff, F. Gotz, A. Otto, J. Westra, and J. M. van Dijl. 2018. A Secreted Bacterial Peptidylarginine Deiminase Can Neutralize Human Innate Immune Defenses. MBio. 9:10.1128/mBio.01704-18. doi: e01704-18 [pii].
121. Gabarrini, G., R. Heida, N. van Ieperen, M. A. Curtis, A. J. van Winkelhoff, and J. M. van Dijl. 2018. Dropping anchor: attachment of peptidylarginine deiminase via A-LPS to secreted outer membrane vesicles of *Porphyromonas gingivalis*. Sci. Rep. 8:8949-018-27223-5. doi: 10.1038/s41598-018-27223-5 [doi].
122. Rawlings, N. D. 2009. A large and accurate collection of peptidase cleavages in the MEROPS database. Database (Oxford). 2009:bap015. doi: 10.1093/database/bap015 [doi].
123. Dalbey, R. E., P. Wang, and J. M. van Dijl. 2012. Membrane proteases in the bacterial protein secretion and quality control pathway. Microbiol. Mol. Biol. Rev. 76:311-330. doi: 10.1128/MMBR.05019-11 [doi].
124. van Roosmalen, M. L., N. Geukens, J. D. Jongbloed, H. Tjalsma, J. Y. Dubois, S. Bron, J. M. van Dijl, and J. Anne. 2004. Type I signal peptidases of Gram-positive bacteria. Biochim. Biophys. Acta. 1694:279-297. doi: S0167488904001235 [pii].
125. Bochtler, M., D. Mizgalska, F. Veillard, M. L. Nowak, J. Houston, P. Veith, E. C. Reynolds, and J. Potempa. 2018. The Bacteroidetes Q-Rule: Pyroglutamate in Signal Peptidase I Substrates. Front. Microbiol. 9:230. doi: 10.3389/fmicb.2018.00230 [doi].
126. Nguyen, M. T., J. Uebele, N. Kumari, H. Nakayama, L. Peter, O. Ticha, A. K. Woischnig, M. Schmaler, N. Khanna, N. Dohmae, B. L. Lee, I. Bekeredjian-Ding, and F. Gotz. 2017. Lipid moieties on lipoproteins of commensal and non-commensal staphylococci induce differential immune responses. Nat. Commun. 8:2246-017-02234-4. doi: 10.1038/s41467-017-02234-4 [doi].
127. Tjalsma, H., A. Bolhuis, J. D. Jongbloed, S. Bron, and J. M. van Dijl. 2000. Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. Microbiol. Mol. Biol. Rev. 64:515-547.
128. Antelmann, H., H. Tjalsma, B. Voigt, S. Ohlmeier, S. Bron, J. M. van Dijl, and M. Hecker. 2001. A proteomic view on genome-based signal peptide predictions. Genome Res. 11:1484-1502. doi: 10.1101/gr.182801 [doi].
129. Tjalsma, H., H. Antelmann, J. D. Jongbloed, P. G. Braun, E. Darmon, R. Dorenbos, J. Y. Dubois, H. Westers, G. Zanen, W. J. Quax, O. P. Kuipers, S. Bron, M. Hecker, and J. M. van Dijl. 2004. Proteomics of protein secretion by *Bacillus subtilis*: separating the "secrets" of the secretome. Microbiol. Mol. Biol. Rev. 68:207-233. doi: 10.1128/MMBR.68.2.207-233.2004 [doi].
130. Sibbald, M. J., A. K. Ziebandt, S. Engelmann, M. Hecker, A. de Jong, H. J. Harmsen, G. C. Raangs, I. Stokroos, J. P. Arends, J. Y. Dubois, and J. M. van Dijl. 2006. Mapping the pathways to staphylococcal pathogenesis by comparative secretomics. Microbiol. Mol. Biol. Rev. 70:755-788. doi: 70/3/755 [pii].
131. Petersen, T. N., S. Brunak, G. von Heijne, and H. Nielsen. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods. 8:785-786. doi: 10.1038/nmeth.1701 [doi].
132. Hiller, K., A. Grote, M. Scheer, R. Munch, and D. Jahn. 2004. PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res. 32:W375-9. doi: 10.1093/nar/gkh378 [doi].
133. Kall, L., A. Krogh, and E. L. Sonnhammer. 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338:1027-1036. doi: 10.1016/j.jmb.2004.03.016 [doi].
134. Juncker, A. S., H. Willenbrock, G. Von Heijne, S. Brunak, H. Nielsen, and A. Krogh. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci. 12:1652-1662. doi: 10.1110/ps.0303703 [doi].
135. Berven, F. S., O. A. Karlsen, A. H. Straume, K. Flikka, J. C. Murrell, A. Fjellbirkeland, J. R. Lillehaug, I. Eidhammer, and H. B. Jensen. 2006. Analysing the outer membrane subproteome of *Methylococcus capsulatus*

(Bath) using proteomics and novel biocomputing tools. Arch. Microbiol. 184:362-377. doi: 10.1007/s00203-005-0055-7 [doi].

136. Lewenza, S., J. L. Gardy, F. S. Brinkman, and R. E. Hancock. 2005. Genome-wide identification of *Pseudomonas aeruginosa* exported proteins using a consensus computational strategy combined with a laboratory-based PhoA fusion screen. Genome Res. 15:321-329. doi: 15/2/321 [pii].

137. Romine, M. F. 2011. Genome-wide protein localization prediction strategies for gram negative bacteria. BMC Genomics. 12 Suppl 1:S1-2164-12-S1-S1. doi: 10.1186/1471-2164-12-S1-S1 [doi].

138. Tjalsma, H., and J. M. van Dijl. 2005. Proteomics-based consensus prediction of protein retention in a bacterial membrane. Proteomics. 5:4472-4482. doi: 10.1002/pmic.200402080 [doi].

139. Gennity, J. M., and M. Inouye. 1991. The protein sequence responsible for lipoprotein membrane localization in *Escherichia coli* exhibits remarkable specificity. J. Biol. Chem. 266:16458-16464.

140. Narita, S., and H. Tokuda. 2007. Amino acids at positions 3 and 4 determine the membrane specificity of *Pseudomonas aeruginosa* lipoproteins. J. Biol. Chem. 282:13372-13378. doi: M611839200 [pii].

141. Schulze, R. J., and W. R. Zuckert. 2006. *Borrelia burgdorferi* lipoproteins are secreted to the outer surface by default. Mol. Microbiol. 59:1473-1484. doi: MMI5039 [pii].

142. Wiker, H. G., M. A. Wilson, and G. K. Schoolnik. 2000. Extracytoplasmic proteins of *Mycobacterium tuberculosis* - mature secreted proteins often start with aspartic acid and proline. Microbiology. 146 (Pt 7):1525-1533.

143. Tjalsma, H., V. P. Kontinen, Z. Pragai, H. Wu, R. Meima, G. Venema, S. Bron, M. Sarvas, and J. M. van Dijl. 1999. The role of lipoprotein processing by signal peptidase II in the Gram-positive eubacterium *Bacillus subtilis*. Signal peptidase II is required for the efficient secretion of alpha-amylase, a non-lipoprotein. J. Biol. Chem. 274:1698-1707.

144. Wang, G., H. Chen, Y. Xia, J. Cui, Z. Gu, Y. Song, Y. Q. Chen, H. Zhang, and W. Chen. 2013. How are the non-classically secreted bacterial proteins released into the extracellular milieu? Curr. Microbiol. 67:688-695. doi: 10.1007/s00284-013-0422-6 [doi].

145. Ebner, P., A. Luqman, S. Reichert, K. Hauf, P. Popella, K. Forchhammer, M. Otto, and F. Gotz. 2017. Non-classical Protein Excretion Is Boosted by PSMalpha-Induced Cell Leakage. Cell. Rep. 20:1278-1286. doi: S2211-1247(17)31024-0 [pii].

146. Krishnappa, L., A. Dreisbach, A. Otto, V. J. Goosens, R. M. Cranenburgh, C. R. Harwood, D. Becher, and J. M. van Dijl. 2013. Extracytoplasmic proteases determining the cleavage and release of secreted proteins, lipoproteins, and membrane proteins in *Bacillus subtilis*. J. Proteome Res. 12:4101-4110. doi: 10.1021/pr400433h [doi].

147. Berks, B. C. 1996. A common export pathway for proteins binding complex redox cofactors? Mol. Microbiol. 22:393-404. doi: 10.1046/j.1365-2958.1996.00114.x [doi].

148. Jeffery, C. J. 1999. Moonlighting proteins. Trends Biochem. Sci. 24:8-11. doi: S0968-0004(98)01335-8 [pii].

149. Ebner, P., J. Rinker, M. T. Nguyen, P. Popella, M. Nega, A. Luqman, B. Schittek, M. Di Marco, S. Stevanovic, and F. Gotz. 2016. Excreted Cytoplasmic Proteins Contribute to Pathogenicity in *Staphylococcus aureus*. Infect. Immun. 84:1672-1681. doi: 10.1128/IAI.00138-16 [doi].

150. Bendtsen, J. D., L. Kiemer, A. Fausboll, and S. Brunak. 2005. Non-classical protein secretion in bacteria. BMC Microbiol. 5:58. doi: 1471-2180-5-58 [pii].

151. Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305:567-580. doi: 10.1006/jmbi.2000.4315 [doi].

152. Berven, F. S., K. Flikka, H. B. Jensen, and I. Eidhammer. 2004. BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res.

32:W394-9. doi: 10.1093/nar/gkh351 [doi].
153. Jones, P., D. Binns, H. Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S. Y. Yong, R. Lopez, and S. Hunter. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236-1240. doi: 10.1093/bioinformatics/btu031 [doi].
154. Finn, R. D., T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H. Y. Chang, Z. Dosztanyi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. Tosatto, C. H. Wu, I. Xenarios, L. S. Yeh, S. Y. Young, and A. L. Mitchell. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 45:D190-D199. doi: 10.1093/nar/gkw1107 [doi].
155. Veith, P. D., N. A. Nor Muhammad, S. G. Dashper, V. A. Likic, D. G. Gorasia, D. Chen, S. J. Byrne, D. V. Catmull, and E. C. Reynolds. 2013. Protein substrates of a novel secretion system are numerous in the Bacteroidetes phylum and have in common a cleavable C-terminal secretion signal, extensive post-translational modification, and cell-surface attachment. J. Proteome Res. 12:4449-4461. doi: 10.1021/pr400487b [doi].
156. Naito, M., H. Hirakawa, A. Yamashita, N. Ohara, M. Shoji, H. Yukitake, K. Nakayama, H. Toh, F. Yoshimura, S. Kuhara, M. Hattori, T. Hayashi, and K. Nakayama. 2008. Determination of the genome sequence of *Porphyromonas gingivalis* strain ATCC 33277 and genomic comparison with strain W83 revealed extensive genome rearrangements in *P. gingivalis*. DNA Res. 15:215-225. doi: 10.1093/dnares/dsn013 [doi].
157. Cock, P. J., J. M. Chilton, B. Gruning, J. E. Johnson, and N. Soranzo. 2015. NCBI BLAST+ integrated into Galaxy. Gigascience. 4:39-015-0080-7. eCollection 2015. doi: 10.1186/s13742-015-0080-7 [doi].
158. Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421-2105-10-421. doi: 10.1186/1471-2105-10-421 [doi].
159. Supek, F., M. Bosnjak, N. Skunca, and T. Smuc. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 6:e21800. doi: 10.1371/journal.pone.0021800 [doi].
160. Baker, S. J., D. J. Payne, R. Rappuoli, E. De Gregorio. 2018. Technologies to address antimicrobial resistance. Proc Natl Acad Sci U S A. 115:12887-12895. doi: 10.1073/pnas.1717160115.
161. Waller, T., L. Kesper, J. Hirschfeld, H. Dommisch, J. Kölpin, J. Oldenburg, J. Uebele, A. Hoerauf, J. Deschner, S. Jepsen, and I. Bekeredjian-Ding. 2016. *Porphyromonas gingivalis* outer membrane vesicles induce selective tumor necrosis factor tolerance in a Toll-Like Receptor 4- and mTOR-dependent manner. Infect. Immun. 84:1194-1204. doi: 10.1128/IAI.01390-15.
162. Fleetwood, A. J., M. K. S. Lee, W. Singleton, A. Achuthan, M. C. Lee, N. M. O'Brien-Simpson, A. D. Cook, A. J. Murphy, S. G. Dashper, E. C. Reynolds, and J. A. Hamilton. 2017. Metabolic remodeling, inflammasome activation, and pyroptosis in macrophages stimulated by *Porphyromonas gingivalis* and its outer membrane vesicles. Front Cell Infect Microbiol. 7:351. doi: 10.3389/fcimb.2017.00351.
163. Frain, K. M., C. Robinson, and J.M. van Dijl. 2019. Transport of Folded Proteins by the Tat System. Protein J. 38:377-388. doi: 10.1007/s10930-019-09859-y.
164. Zhao X., L. M. Palma Medina, T. Stobernack, C. Glasner, A. de Jong, P. Utari, R. Setroikromo, W. J. Quax, A. Otto, D. Becher, G. Buist, and J. M. van Dijl. 2019. Exoproteome heterogeneity among closely related *Staphylococcus aureus* t437 isolates and possible implications for virulence. J Proteome Res. 18:2859-2874. doi: 10.1021/acs.jproteome.9b00179.
165. Jeong, S. H., Y. Nam, H. Jung, J. Kim, Y. A. Rim, N. Park, K. Lee, S. Choi, Y. Jang, Y. Kim, J. H. Moon, S. M. Jung, S. H. Park, and J. H. Ju. 2018. Interrupting oral infection of Porphyromonas gingivalis with anti-FimA antibody attenuates bacterial dissemination to the arthritic joint and improves experimental arthritis. Exp

Mol Med. 50:e460. doi: 10.1038/emm.2017.301.

CHAPTER 4

SIGNATURES OF CYTOPLASMIC PROTEINS IN THE EXOPROTEOME DISTINGUISH COMMUNITY- AND HOSPITAL-ASSOCIATED METHICILLIN-RESISTANT *STAPHYLOCOCCUS AUREUS* USA300 LINEAGES

Solomon A. Mekonnen, Laura M. Palma Medina, Corinna Glasner, Eleni Tsompanidou, Anne de Jong, Stefano Grasso, Marc Schaffer, Ulrike Mäder, Anders R. Larsen, Heidi Gumpert, Henrik Westh, Uwe Völker, Andreas Otto, Dörte Becher, Jan Maarten van Dijl

Virulence, 2017, 8(6): 891-907

Abstract

Methicillin-resistant *Staphylococcus aureus* (MRSA) is the common name for a heterogeneous group of highly drug-resistant staphylococci. Two major MRSA classes are distinguished based on epidemiology, namely community-associated (CA) and hospital-associated (HA) MRSA. Notably, the distinction of CA- and HA-MRSA based on molecular traits remains difficult due to the high genomic plasticity of *S. aureus*. Here we sought to pinpoint global distinguishing features of CA- and HA-MRSA through a comparative genome and proteome analysis of the notorious MRSA lineage USA300. We show for the first time that CA- and HA-MRSA isolates can be distinguished by two distinct extracellular protein abundance clusters that are predictive not only for epidemiologic behavior, but also for their growth and survival within epithelial cells. This 'exoproteome profiling' also groups more distantly related HA-MRSA isolates into the HA exoproteome cluster. Comparative genome analysis suggests that these distinctive features of CA- and HA-MRSA isolates relate predominantly to the accessory genome. Intriguingly, the identified exoproteome clusters differ in the relative abundance of typical cytoplasmic proteins, suggesting that signatures of cytoplasmic proteins in the exoproteome represent a new distinguishing feature of CA- and HA-MRSA. Our comparative genome and proteome analysis focuses attention on potentially distinctive roles of 'liberated' cytoplasmic proteins in the epidemiology and intracellular survival of CA- and HA-MRSA isolates. Such extracellular cytoplasmic proteins were recently invoked in staphylococcal virulence, but their implication in the epidemiology of MRSA is unprecedented.

Keywords: community, epithelial cells, exoproteome, hospital, moonlighting, MRSA, protein secretion, Staphylococcus, USA300, virulence factor

Supplementary files avaialble at: https://github.com/grassoste/Thesis-supplementary-files

Introduction

Staphylococcus aureus is a wide-spread commensal bacterium, but also a notoriously drug-resistant pathogen that causes a wide range of diseases, varying from mild skin infections to life-threatening invasive diseases [1]. About 20-30 % of the healthy human population is known to carry *S. aureus*, the anterior nares being the preferred niche [2].

Since the clinical implementation of antibiotics, *S. aureus* has acquired a range of resistance traits through mutations and horizontal gene transfer. This has culminated in the emergence of methicillin-resistant *S. aureus* (MRSA), a major healthcare problem world-wide [3,4]. The emergence of MRSA is a particularly worrisome development since it is associated with increased morbidity and mortality, especially if very young, immune-compromised or elderly individuals are infected [5,6]. Moreover, no effective vaccine against MRSA is currently available [7–9].

Two major classes of MRSA are currently distinguished based on their epidemiology, namely community-associated (CA) and hospital-associated (HA) MRSA. CA-MRSA is mainly a threat to healthy individuals, causing in particular skin and soft tissues infections, but also serious invasive infections such as pneumonia and osteomyelitis [10–13]. In contrast, HA-MRSA infections are associated with prolonged hospitalization, stay in intensive care units, hemodialysis, surgery, and long-term exposure to antibiotics [14].

Molecular markers for high-confidence distinction between CA- and HA-MRSA isolates are urgently needed in the prevention and control of hospital outbreaks. Different DNA typing methods, such as pulsed-field gel electrophoresis (PFGE) and *Staphylococcus* protein A (*spa*) typing have been used to differentiate between these two classes of MRSA [15]. This was so far feasible, because particular *S. aureus* lineages with distinct sequence types are associated with the CA- or HA-associated behavior. In addition, particular virulence genes (e.g. for the Panton-Valentin leukocidin; PVL), the arginine catabolic mobile element (ACME), and mobile genetic elements carrying the *mecA* gene for methicillin resistance are used to distinguish CA- and HA-MRSA [11–14,16,17]. However, such DNA-based typing methods do not allow easy distinction between closely related CA- and HA-MRSA lineages, because the causative molecular features have remained largely enigmatic. For instance, PFGE assigns CA-MRSA isolates with the *spa* type t008 and HA-MRSA isolates with the *spa* type t024 to the same USA300 lineage [18]. Likewise, *spa* typing has insufficient discriminatory power to distinguish closely related CA and HA isolates as it assigns CA-USA300 isolates with the multi-locus sequence type ST8 and more distantly related HA isolates with the sequence type ST8 to the same *spa* type t008 [18]. Nevertheless, we have previously shown that a multiple-locus variable number tandem repeat fingerprinting (MLVF) approach may distinguish these highly related *S. aureus* isolates [19].

An important challenge for the clinic is that *S. aureus* types previously regarded as CA, such as USA300 and the European ST80 clone, are becoming common hospital pathogens causing outbreaks [18,20–22]. Clearly, an increasing prevalence of CA-MRSA in the community makes it harder to exclude the respective lineages from hospitals, because they can be carried into the hospitals by MRSA-positive patients, healthcare workers and visitors. Furthermore, it is conceivable that these bacteria have acquired, either before or after entry into the hospital environment, properties that facilitate their spread in this setting. The latter view would be supported by the observation that the closely related USA300 isolates with *spa* types t008 and t024 display different epidemiology [18].

The distinction of CA- and HA-MRSA at the molecular level is challenging, because many factors may contribute to bacterial epidemiological behavior, not in the last place interactions with the human host. High-throughput analytical 'omics' approaches, especially genomics and proteomics, are particularly suitable for exploring such multi-factorial behavior since they allow the definition of feature- or condition-specific signatures [23,24]. Furthermore, proteomics applied to bacterial pathogens grown under infection-mimicking conditions is a powerful tool for investigating different lineage- or type-specific patterns of gene expression

[25]. In the context of infection-related research, it is important to focus special attention on the extracellular proteome ('exoproteome') as it represents the main reservoir of virulence factors that are first in interacting with the human host [26,27]. Specifically, secreted toxins and other virulence factors of *S. aureus* contribute to tissue damage, host invasion, and evasion of the host's immune responses [28,29]. Thus, proteomics has a high potential for identifying diagnostic biomarkers, and novel vaccine or drug targets [30].

To obtain a better understanding of the molecular differences between CA- and HA-MRSA, the present study was aimed at a global comparative genome and exoproteome analysis of 12 MRSA isolates belonging to the USA300 lineage as defined by PFGE. As these isolates were all collected from Denmark (DK), we refer to them as the CA^{DK} and HA^{DK} isolates. Specifically, the CA^{DK} group had the sequence type ST8, the *spa* type t008 and was PVL-positive, whereas the HA^{DK} group was characterized by the sequence type ST8 and the *spa* type t024 [18,21]. As a control group, we also investigated the exoproteomes of three HA-MRSA isolates from the Dutch (NL) - German (DE) border region, here referred to as HA^{NL-DE}, which have the sequence type ST8, and *spa* type t008 or t024 [19]. The genomes of all 15 isolates were sequenced, and their extracellular proteins were analyzed by liquid chromatography and mass spectrometry (LC-MS). In brief, CA and HA isolates could be distinguished to some extent by the accessory genome. More importantly, a principal component analysis (PCA) of the exoproteome MS data clustered the 15 investigated isolates into two groups that match their different epidemiological behavior.

Results

Comparative genomic analysis

Whole genome sequence analysis was performed to determine the genomic similarities and differences of all 15 investigated isolates. A phylogenetic tree based on the core genome of the isolates showed that the six CA^{DK}, and five of the six HA^{DK} isolates formed two distinct clusters (Fig. 1). One HA^{DK} isolate (D3) showed a more distant relationship with the other HA^{DK} isolates. Furthermore, the three HA^{NL-DE} isolates formed a separate cluster that is closer to the CA^{DK} than the HA^{DK} isolates. In addition to the phylogenetic analysis, a comparative analysis of the accessory genomes of the isolates was performed, which is presented as a heatmap in Figure 1. As illustrated in the heat map, the CA isolates have overall more accessory genes than the HA isolates. Perhaps more importantly, the clustering of accessory genes is indicative of a separation between the CA and HA isolates, irrespective of the geographical origin of the HA isolates. This separation is also reflected in the presence or absence of a number of known virulence genes (red lines in Fig. 1), such as the PVL-encoding genes *lukF* and *lukS* that were exclusively found in the CA isolates, and the enterotoxin-encoding genes *sea*, *sed*, *sej*, and *ser* that were only present in the investigated HA isolates (Supplementary Table 1). Of note, PVL is often used as a marker for CA-MRSA and enterotoxin genes appear to be rare in CA isolates of the USA300 lineage [31], but a possible association of enterotoxin genes with HA behavior would be novel.

Both CA- and HA-MRSA isolates carried a *norA* gene that provides resistance to fluoroquinolones, and *mecA* and *blaZ* genes for β-lactam resistance (Supplementary Table 2). Genes potentially providing resistance to macrolides, lincosamides and streptogramin B (*msr*(*A*)), aminoglycosides (*aph*(3')-III), and macrolides (*mph*(*C*)) were exclusively identified in the CA-MRSA isolates, whereas *erm*(*A*) and *spc* that provide resistance to macrolides and aminoglycosides, respectively, were exclusively identified among the HA-MRSA isolates (Supplementary Table 2). Altogether, the CA-MRSA isolates carried more (potential) antimicrobial resistance genes than the investigated HA-MRSA isolates.

Unique and shared exoproteins

To characterize the exoproteomes of the 15 MRSA isolates, they were cultured in RPMI medium since a recent study showed that global gene expression profiles of *S. aureus* cells grown in RPMI or human plasma

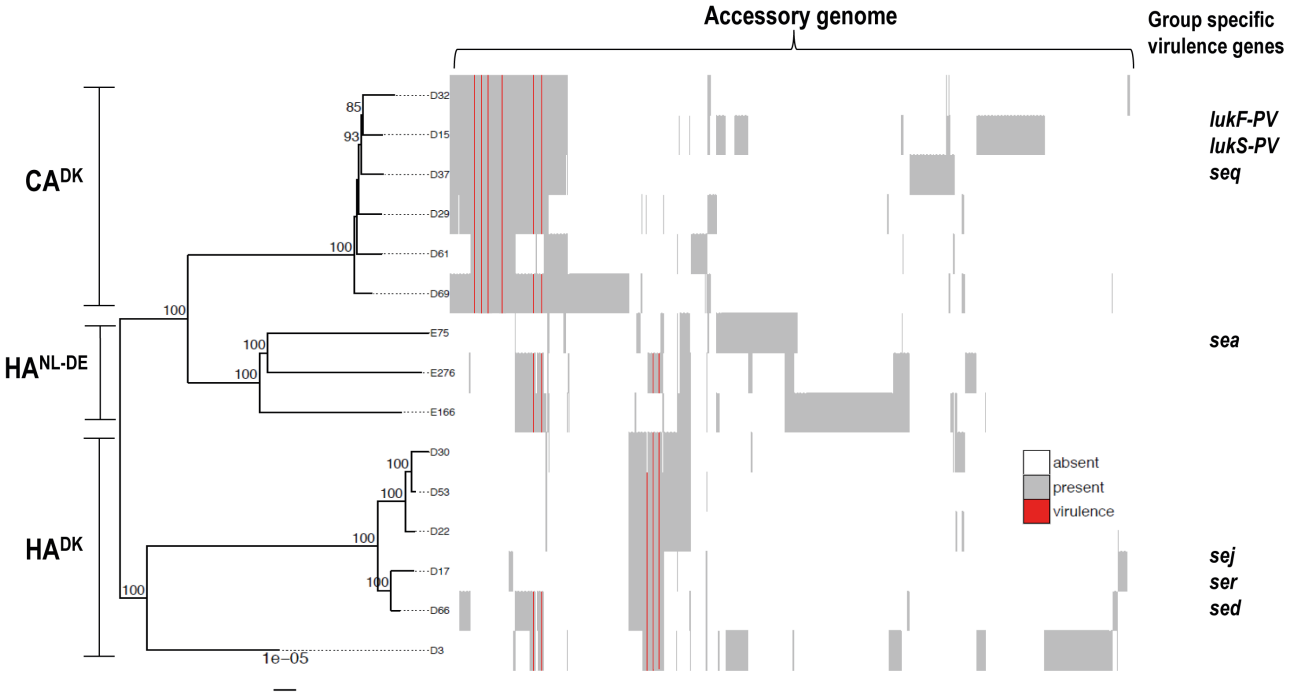


Figure 1. Phylogenetic tree and accessory genomes of all 15 investigated CA^{DK}, HA^{DK} and HA^{NL-DE} isolates. The tree is midpoint rooted and bootstrap support >70% is indicated on the branches. The heatmap to the right of the phylogenetic tree illustrates the accessory genome. The columns of the heatmap are hierarchically clustered based on the presence/ absence of genes. Known virulence genes are indicated in red. Examples of virulence genes that are exclusively present in one of the three groups are indicated as group-specific virulence genes.

are highly similar [24]. Samples were withdrawn for exoproteome analyses at mid-exponential growth phase and 90 min after entry into the stationary phase. No major differences in the growth curves of the 15 MRSA were observed (data not shown). As shown by gel-free mass spectrometry, a total number of 409 unique proteins was identified from the 15 exoproteome samples of exponentially grown isolates. Similarly, a total number of 458 unique proteins was identified from the 15 exoproteome samples generated from stationary phase cultures. Proteins were considered for further analyses when they were present in at least 50 % of the isolates of a particular group, i.e. when a protein was present in three out of the six isolates in CA^{DK} and HA^{DK}, and in two out of three isolates in HA^{NL-DE}. Thus, 283 and 307 unique proteins identified in the exponential or stationary phase samples, respectively, were included in the subsequent analyses (Supplementary Table 3). The majority of these proteins was shared by all three groups both in the exponential (Fig. 2a) and stationary (Fig. 2b) growth phases. Importantly, there are more proteins shared by the HA^{DK} and HA^{NL-DE} isolates than by the HA^{DK} or HA^{NL-DE} isolates and CA^{DK} isolates. This implies that, in terms of exoprotein production, the two groups of HA isolates are more closely related with each other than the HA and CA isolates. Furthermore, unique proteins ranging from 2 to 12 proteins in the exponential growth phase, and 3 to 10 proteins from the stationary growth phase, which were specific to only one of the three groups of isolates were identified (Fig. 2a,b). Together, these data show that the majority of extracellular proteins of the CA^{DK}, HA^{DK} and HA^{NL-DE} is common. Yet, a subset of the exoproteins appears to be specific for each of the three groups of isolates.

Predicted sub-cellular localization of identified exoproteins

Bacterial exoproteomes are known to contain proteins that are actively secreted and proteins that are liberated from the cells through (auto-)lysis or other unidentified 'non-classical secretion' mechanisms [27,32–37].

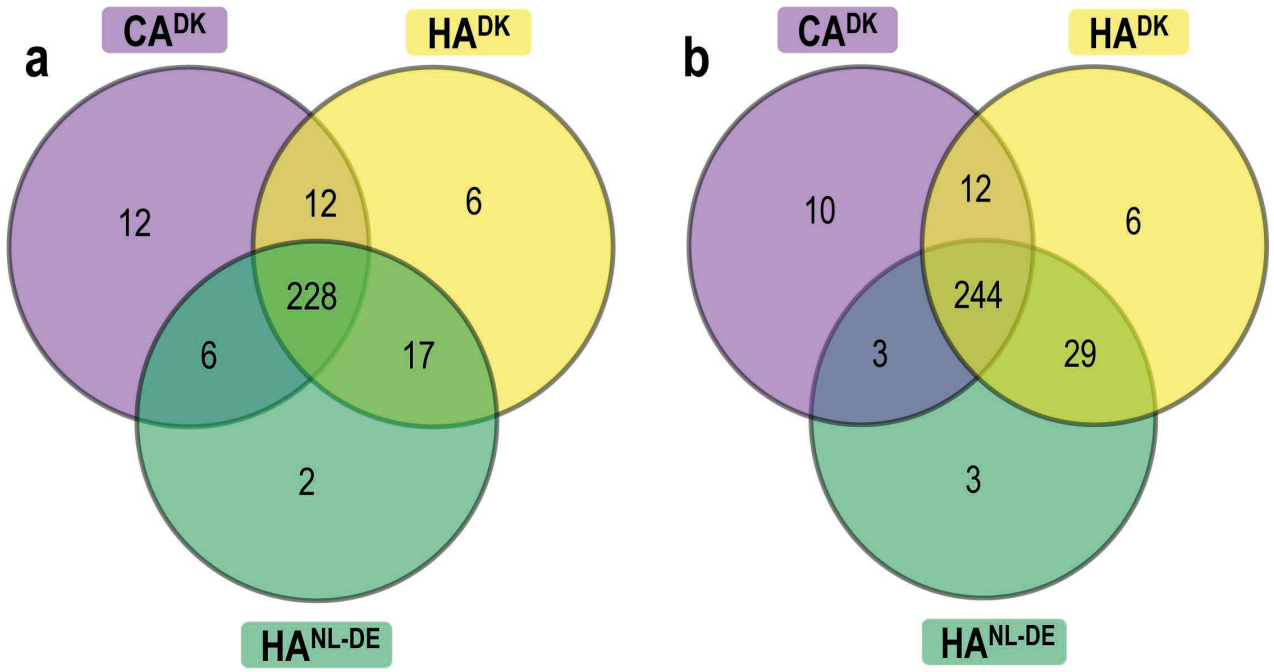


Figure 2. Shared and uniquely identified proteins in CA^{DK}, HA^{DK} and HA^{NL-DE} *S. aureus* isolates. The Venn diagrams relate to cells in the exponential (**a**) and stationary (**b**) growth phases. The numbers of commonly and uniquely identified proteins of the different groups of isolates are indicated.

These proteins can be distinguished through signal peptide predictions, which is relevant as most known virulence factors contain signal peptides to direct their export from the cytoplasm [26]. Thus, we predicted the sub-cellular localization of proteins that were identified by MS. The vast majority of the proteins identified in the exoproteomes of the isolates in the exponential and stationary growth phases were assigned to the class of cytoplasmic proteins followed by secreted proteins, lipoproteins, cytoplasmic membrane proteins and cell wall-associated proteins (Fig. 3a,b). Notably, in the exponential growth phase, the numbers of accessory exoproteins that were predicted as cytoplasmic were higher in the CA^{DK} group than in the HA^{DK} and HA^{NL-DE} groups (Fig. 3c). Conversely, in the stationary phase, the numbers of accessory exoproteins predicted as cytoplasmic were higher among the HA^{DK} and HA^{NL-DE} groups than in the CA^{DK} group (Fig. 3d). For exoproteins with a predicted localization in the membrane (i.e. membrane- and lipoproteins) or cell wall no major differences were observed in the three groups, irrespective of the growth phase (data not shown). Lastly, higher numbers of predicted secretory proteins were identified in growth media of the HA group than the CA group in both growth phases. Altogether, these data imply that the investigated CA and HA isolates are similar in terms of the predicted localization of their exoproteins. Nonetheless, the main distinction among these groups was the time point at which cytoplasmic proteins are liberated from the cells.

Relative extracellular abundance of known and putative virulence factors

To obtain more comprehensive insights in the possible differences in the levels of known extracellular virulence factors, we assessed their relative abundance for the different investigated isolates. Detailed evaluation of the normalized spectral counts showed differential and similar expression levels for 24 virulence factors among the three groups of isolates both in the exponential growth phase (Fig. 4a), and in the stationary growth phase (Fig. 4b). Of note, neither PVL nor enterotoxins that were identified as potentially distinguishing features for CA and HA isolates based on the genome sequence were detectable in the extracellular proteome. On the other hand, statistically significantly different levels of the IsdA, IsdB, SCIN and Vwb proteins were identified in the growth media of exponentially growing isolates, and the same was true for the Ebps, IsdB, SCIN and

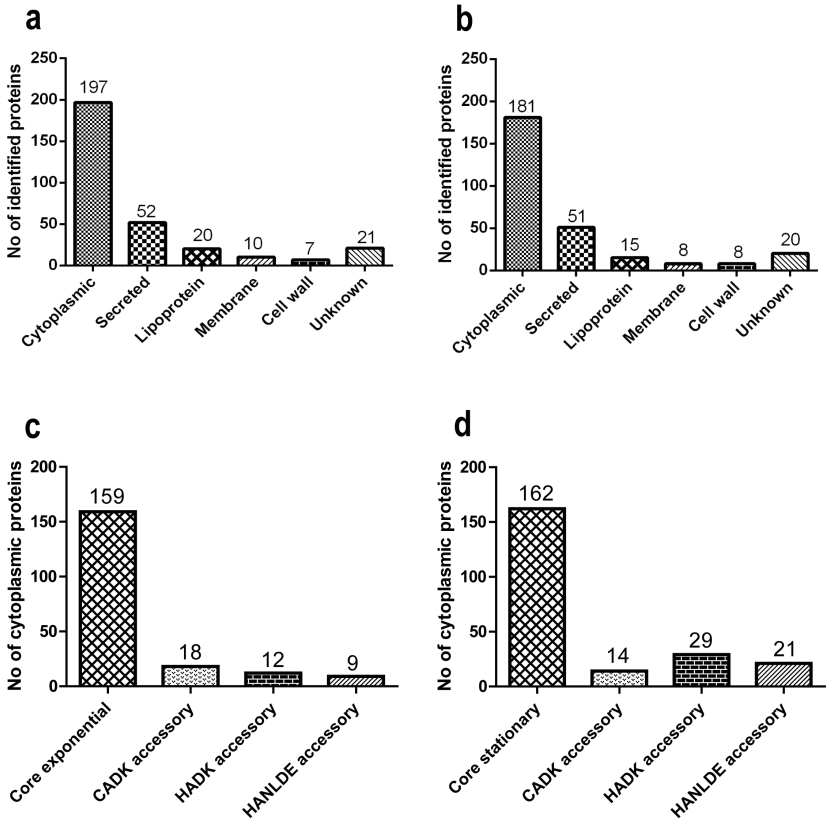


Figure 3. Predicted subcellular localization of identified extracellular proteins. The predicted subcellular localization of all 494 identified extracellular proteins is shown for cells in the exponential (**a**) and stationary (**b**) growth phases. Panels (**c**) and (**d**), respectively, highlight the appearance of predicted cytoplasmic core and accessory cytoplasmic proteins in the growth medium of the CA^{DK}, HA^{DK} and $HA^{NL\text{-}DE}$ isolates in the exponential and stationary growth phases. The numbers of proteins identified in each category are indicated at the top of the bars.

Vwb proteins in the growth media of stationary growing isolates (Fig. 4a,b, Supplementary Table 4).

The relative amounts of individual secreted exoproteins are likely important for the behavior of the respective *S. aureus* isolate, especially where this concerns secreted toxins or immune evasion factors. Therefore, we determined the relative abundance of proteins in the three groups of isolates from the normalized spectral counts of proteins. A volcano plot was used to present the proteins that were detectable at statistically significantly higher or lower levels among the three groups of isolates during both the exponential and stationary growth phases. From the total of 283 proteins identified in samples collected in the exponential growth phase, a relatively large number of proteins was present at statistically significantly different levels when the CA^{DK} and the two HA isolate groups (*i.e* HA^{DK} and $HA^{NL\text{-}DE}$) were compared, and this difference was larger than the difference between the HA^{DK} and $HA^{NL\text{-}DE}$ isolates (Fig. 5a; Supplementary Table 4). A similar pattern was observed for the samples harvested during the stationary phase (Fig. 5b). Additionally, some proteins were exclusively present in one group of isolates, e.g. the chemotaxis inhibitory protein (CHIPS) was identified only in CA^{DK}, the enterotoxin type D only in HA^{DK}, and the enterotoxin type A only in HA^{NLDE} isolates, and this applied both to exponential and stationary phase growth medium samples (Supplementary Table 5). Together, these data show differences in the relative abundance of extracellular proteins at statistically significant levels in all the three groups of isolates, but especially for the CA and HA isolates.

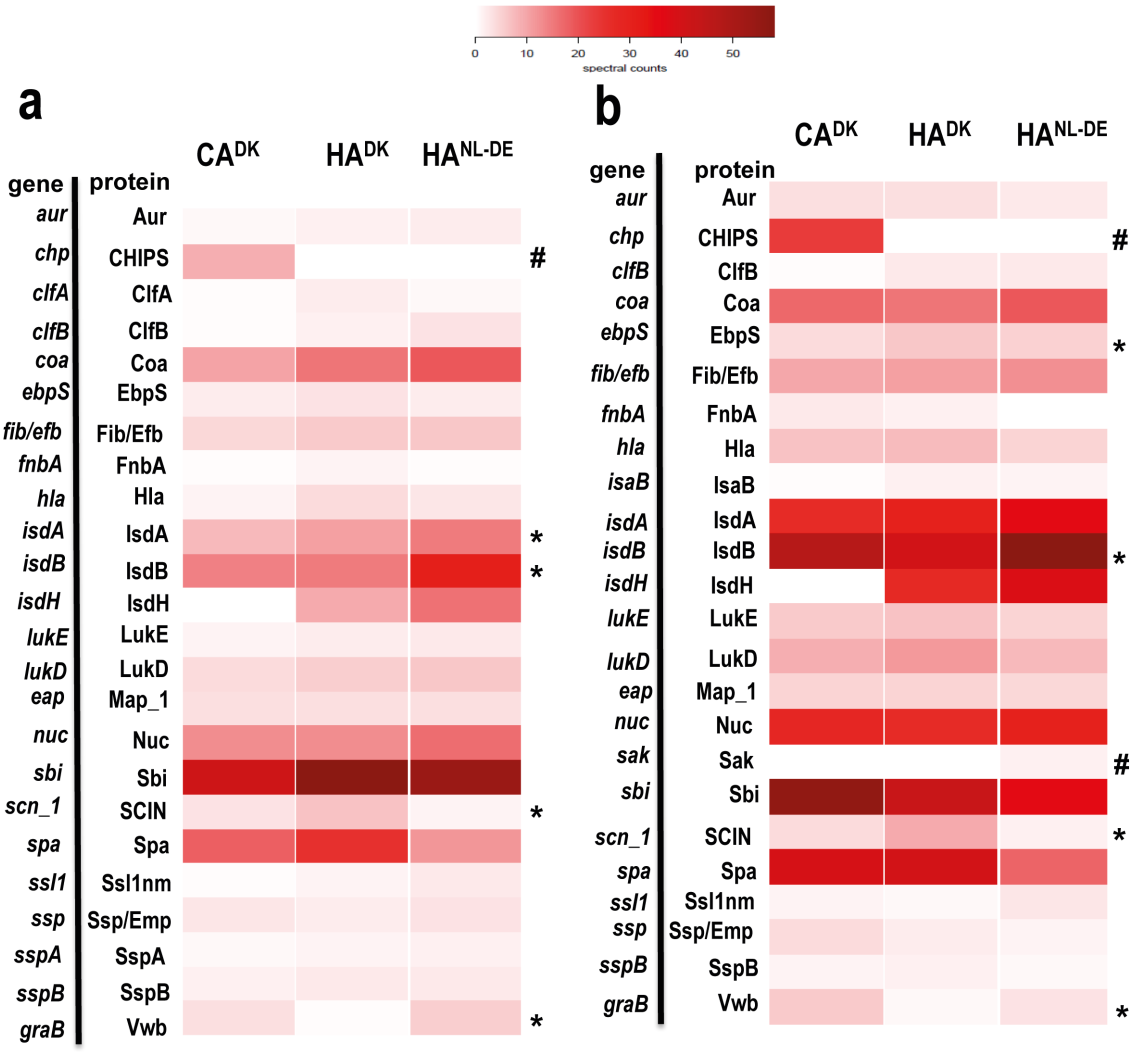


Figure 4. Heat map analysis of quantified extracellular virulence factors. The normalized spectral counts of known extracellular virulence factors identified by Mass Spectrometry in growth media of the three groups of isolates are graphically represented as colored heat maps. Each heat map includes three columns representing each of the three groups of the isolates. Of note, each column of CA^{DK} and HA^{DK} isolates is based on the average of six different isolates each analyzed in duplicate, and the HA^{NL-DE} column is based on the average of three different isolates each analyzed in duplicate. Each row represents a particular protein. Panels (**a**) and (**b**) represent known virulence factors of *S. aureus* as identified in the growth medium fractions of cells in the exponential and stationary growth phases, respectively. *Statistically significant differences in relative abundance of the proteins marked between the groups; $^{\#}$Proteins present in one group of isolates only.

Levels of mRNA for selected exoproteins

The abundance of a bacterial exoprotein reflects the net result of transcription of the respective gene, mRNA translation, translocation of the precursor protein across the membrane, post-translocational folding of the protein into a stable conformation, cell wall passage and the protein's stability in the bacterial extracellular milieu. This implies that extracellular protein abundance is not always linearly correlated with the transcript levels. Yet, mRNA levels are major determinants for protein expression levels. Therefore, a Northern blotting analysis was performed to assess whether there is a possible correlation between transcript levels and extracellular protein abundance. Specifically, we compared the transcript levels for a secreted virulence factor (*ebpS*) and two cytosolic proteins (*fabF* and *rpoB*). Consistent with the MS data, in the stationary phase, the

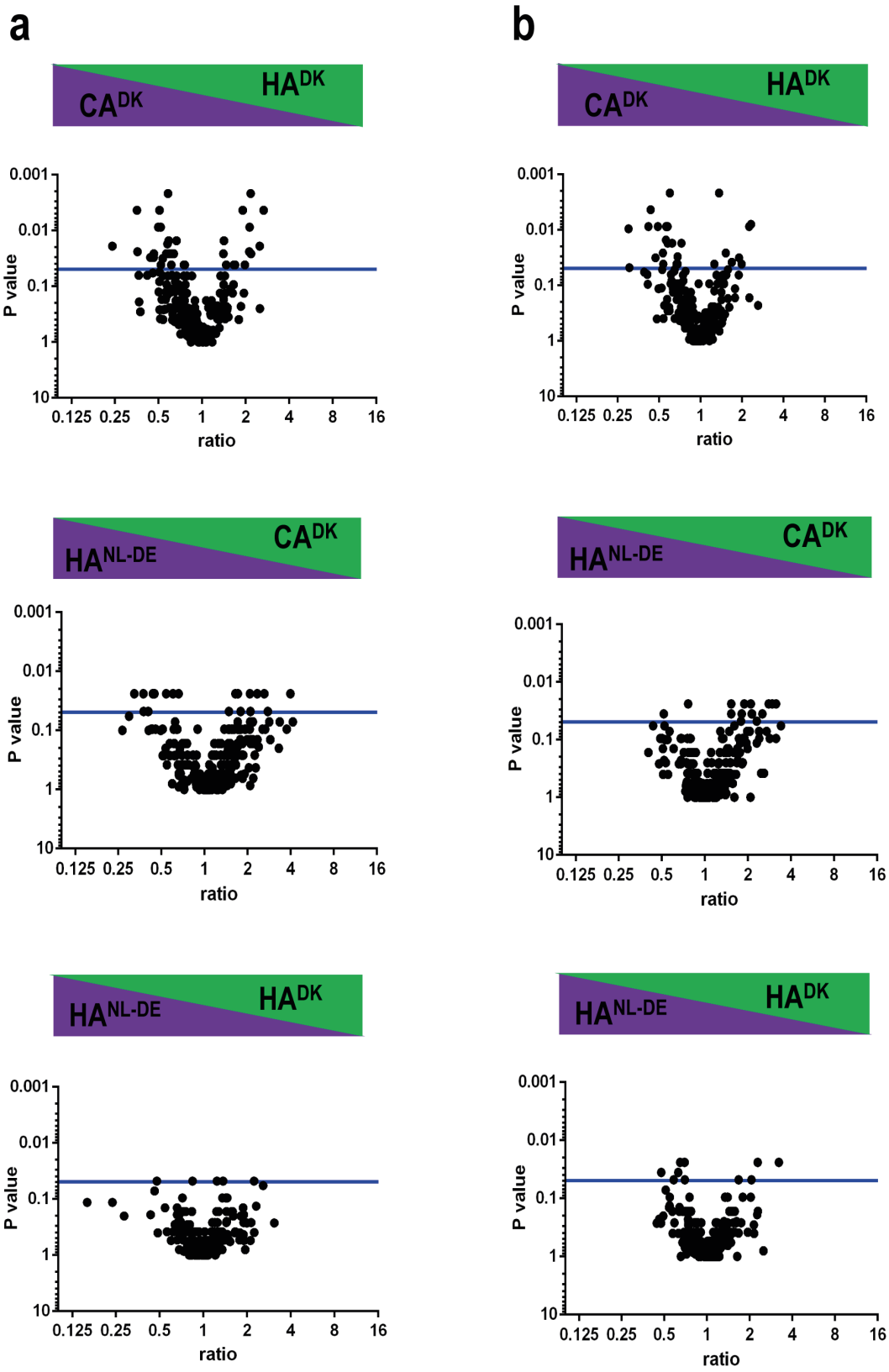


Figure 5. Differences in relative extracellular protein abundance in CA^{DK}, HA^{DK} and HA^{NL-DE} isolates. Statistically significant differences in the relative abundance of identified extracellular proteins are presented in volcano plots for samples collected in the exponential (**a**) and stationary (**b**) growth phases. Horizontal blue lines indicate a p-value threshold of 0.05.

mRNA level of *ebpS* was higher in HA isolates than in the CA isolates (Fig. 6a), whereas the mRNA levels of *fabF* and *rpoB* were higher in the CA isolates compared to the HA isolates (Fig. 6b,c). Of note, the *fabF* and *rpoB* mRNA levels in the CA^{DK} isolate D29 were more similar to the respective mRNA levels in HA^{DK} isolates than to those in the CA^{DK} isolates. On the other hand, two of HA^{DK} isolates (D30 and D66), displayed *fabF* and *rpoB* mRNA levels comparable to those observed for the CA^{DK} isolates.

Northern blotting analyses can also provide information on the expression of genes for which the encoded proteins were not covered by the proteome analysis. Since phenol-soluble modulins (PSMs) are particularly relevant for virulence, but notoriously difficult to identify by proteomics due to their small size, we investigated the *psmα1-4* mRNA levels by Northern blotting. In the stationary growth phase, the *psmα1-4* mRNA levels were higher in most of the HA^{DK} isolates than in the CA^{DK} isolates (Fig. 6d). However, also the CA^{DK} isolate D29 showed a relatively high level of *psmα1*-4 mRNA that was comparable to the *psmα1*-4 mRNA levels in the HA^{DK} isolates (Fig. 6d). Based on these Northern blotting data, the proteomics data were reassessed with less stringent criteria where we considered also proteins identified with only one peptide. Thus, we were able to identify both the PSMβ1 and PSMβ2 proteins in medium fractions of five out of the six HA^{DK} isolates grown to stationary phase. Of note, the same was true for the D29 CA^{DK} isolate. These findings

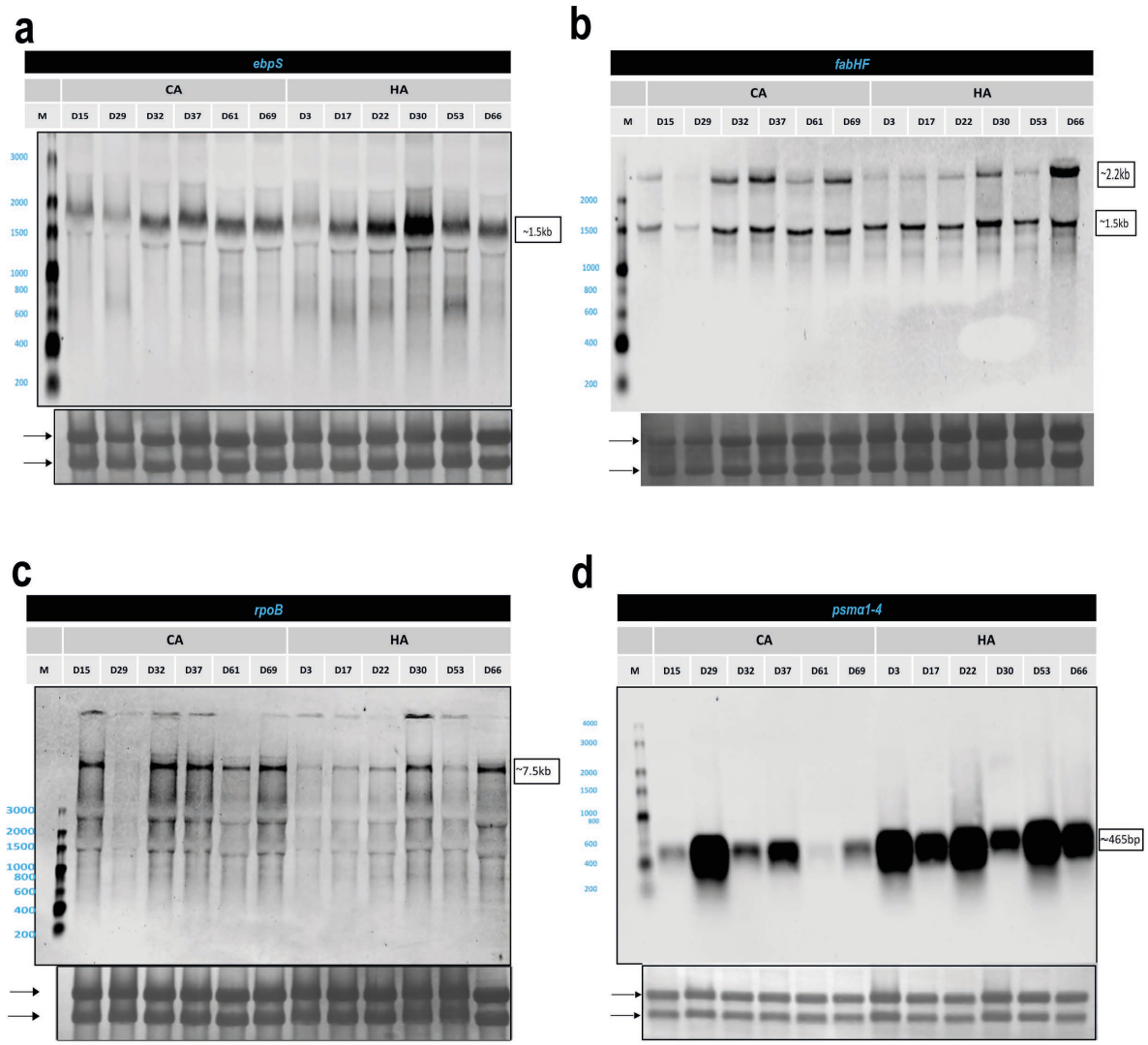


Figure 6. Northern blotting analysis of selected genes. Arrows mark the positions of 23S- and 16S-rRNA bands on the methylene blue stained membranes. Sizes of specific transcripts are indicated on the right side of each display. In case of *fabHF* two bands were detected, the larger band of ~2.2 kb representing a *fabHF* transcript and the lower band of ~1.5 kb only *fabF*. (**a**) *ebps*, (**b**) *fabHF*, (**c**) *rpoB*, and (**d**) *psmα1*-4.

are fully consistent with the relative mRNA levels detected by Northern blotting.

Clustering of CA and HA isolates based on exoproteome abundance signatures

Principal component analysis (PCA) was performed to assess the overall relationships between the different investigated isolates in terms of their exoproteome profiles. Of note, this PCA was based on the normalized spectral counts of proteins that were produced by all three groups of isolates, specifically 283 proteins from exponentially growing bacteria, and 308 proteins from bacteria in the stationary growth phase. Importantly, the PCA analysis revealed that the CA- and HA-MRSA isolates clustered in two distinct groups based on the 'exoprotein abundance signatures' where the HA cluster included both the $HA^{NL\text{-}DE}$ and the HA^{DK} isolates (Fig. 7a,b) irrespective of their geographical origin. Yet, the HA cluster included two CA^{DK} isolates (D29 and D61), whose exoprotein abundance signatures apparently resemble those of the analyzed HA isolates.

The PCA analysis in Figure 7 was based on all identified extracellular proteins, including proteins with different predicted subcellular localizations. To assess whether the discriminating information relates to proteins with a particular predicted localization site, PCA analysis was performed on: i, predicted cytoplasmic proteins alone, ii, all identified proteins except the predicted cytoplasmic proteins, and iii, all identified proteins except the predicted cytoplasmic proteins and the proteins of unknown localization. Unexpectedly, the distinguishing information was primarily associated with the predicted cytoplasmic proteins (Supplementary Table 6).

Voronoi treemaps can be applied to link quantitative proteomic data and functional classifications. Thus, we used Voronoi treemaps to characterize the extracellular proteins identified for the three groups of isolates. The biological functions of the identified proteins were mainly related to protein biosynthesis, carbohydrate and carbon metabolism, oxidative stress, and adhesion (Supplementary Figure 1a,b). Of note, adhesion-associated extracellular proteins were somewhat more pronounced among the HA- than the CA isolates in the exponential growth phase (Supplementary Figure 1a). Conversely, adhesion-related extracellular proteins were more prominently present in CA- than in HA isolates in the stationary growth phase (Supplementary Figure 1b). Despite the fact that there were unique proteins identified in each of the

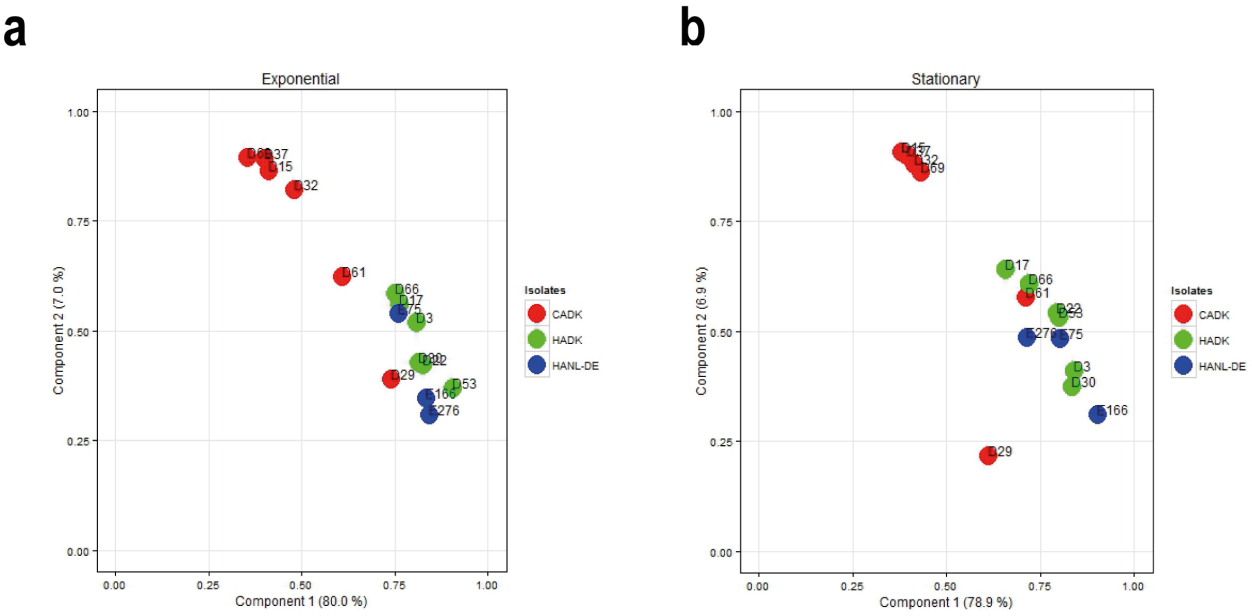


Figure 7. Principal component analysis (PCA) of the normalized spectral counts of identified extracellular proteins. Two-dimensional PCA plots are displayed for growth medium samples from the 15 CA^{DK}, HA^{DK} and $HA^{NL\text{-}DE}$ isolates in the (**a**) exponential growth phase and (**b**) stationary growth phase.

three groups of isolates, no major differences in the overall functions of the identified extracellular proteins were observed.

Differences in staphylococcal survival within epithelial cells

Since the PCA analysis of exoproteins grouped the studied *S. aureus* isolates into two distinct clusters, we asked the question whether these groups might interact differently with human host cells. A bronchial epithelial cell line (16HBE14o-) was selected for this purpose, because CA-MRSA isolates have been implicated in severe respiratory infections among healthy individuals from the community. This fact might indicate a superior ability of this group of bacteria to interact with airway cells. Furthermore, the 16HBE14o- epithelial cell line forms a confluent layer that allows the monitoring of infecting bacteria over several days. Thus, *in vitro* cultured 16HBE14o- epithelial cells were infected with the CA^{DK}, HA^{DK} and $HA^{NL\text{-}DE}$ isolates, and the subsequent binding, internalization and intracellular survival of staphylococci were assessed through counting by flow cytometry. While the staphylococcal isolates did not show major differences in the internalization rate into the epithelial cells, there was a marked difference in post-internalization growth and survival. As shown in Figure 8 and Supplementary Figure 2, four CA^{DK} isolates (D15, D32, D37, D69) were able to multiply inside the epithelial cells during the first two days post infection, after which the population size decreased. In contrast, the HA^{DK} isolates did not multiply with the exception of isolate D17, which showed a slight increase, comparable to the CA^{DK} isolate D69 (Fig. 8). Of note, the CA^{DK} isolates D29 and D61, which were grouped with the HA^{DK} isolates in the exoproteome PCA analysis (Fig. 7), displayed similar intracellular behavior as the five HA^{DK} isolates D03, D22, D30, D53 and D66. Notably, all the three $HA^{NL\text{-}DE}$ isolates showed a similar intracellular behavior as the HA^{DK} isolates. These findings imply that the distinction of the investigated HA and CA isolates based on differences in their exoproteomes, is reflected in their growth and survival behavior upon internalization of

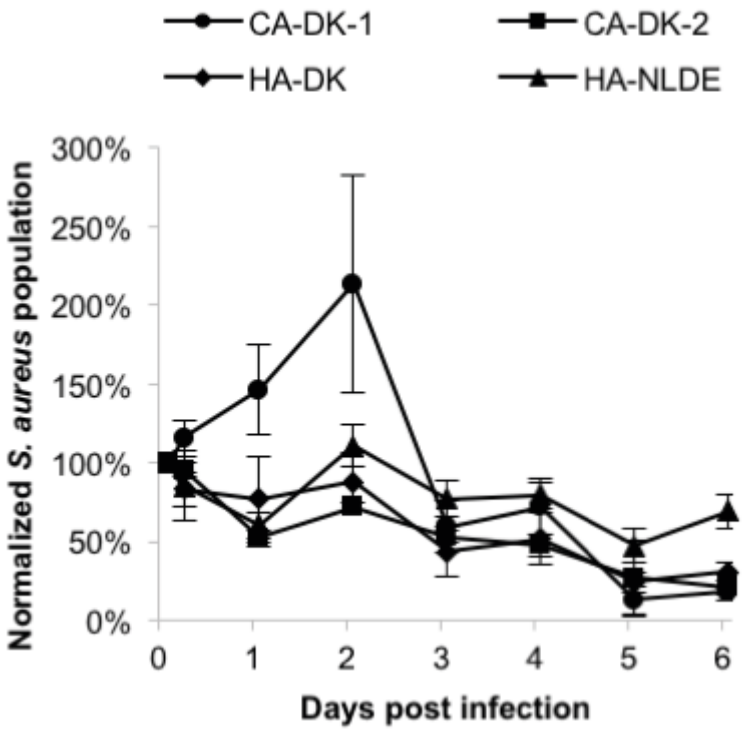


Figure 8. Survival of CA and HA isolates internalized by 16HBE14o- bronchial epithelial cells. Averaged survival curves are shown for the CA^{DK}, HA^{DK} and $HA^{NL\text{-}DE}$ isolates, where the CA^{DK} isolates are separated into two groups in accordance with their exoprotein abundance signatures in **Fig. 8** (CA-DK-1 includes isolates D15, D32, D37 and D69; CA-DK-2 includes isolates D29 and D61).

human bronchial epithelial cells.

Discussion

The serious threat that MRSA represents for hospitalized patients demands an accurate and reliable method of distinction between CA- and HA-MRSA isolates for the purpose of prevention and control of outbreaks. In the present study, we explored the feasibility of applying a combined genome and proteomics-based approach to distinguish MRSA isolates with different epidemiological behavior. To facilitate the interpretation of the complex data, we used a set of genetically closely related CA- and HA-MRSA isolates with the sequence type ST8. Altogether, our findings revealed several distinguishing features between the investigated CA and HA isolates at the levels of the accessory genome and the exoproteome.

Since exoproteins have major roles in staphylococcal colonization of the host and virulence, we compared the exoproteome profiles of three genetically similar, but epidemiologically unrelated groups of MRSA isolates, i.e. the CA^{DK}, HA^{DK} and HA^{NL-DE} isolates. Noticeably, the vast majority of identified exoproteins was shared by all the three groups of isolates. However, some of the identified exoproteins were unique for a particular group of isolates, while the abundance of several common proteins varied among the three groups of isolates. This probably reflects the fact that the *S. aureus* exoproteome is heterogeneous due to this organism's genomic plasticity [26,27]. Nonetheless, the investigated HA^{NL-DE} isolates shared more similarities with respect to accessory virulence genes and actually detected exoproteins with the HA^{DK} isolates than with the CA^{DK} isolates, even though their core genome was more closely related to that of the CA^{DK} isolates. The latter finding is in line with previously reported observations that genetically closely related *S. aureus* isolates may reveal heterogeneous exoproteome profiles [38,39]. Taken together, our combined observations imply that both qualitative and quantitative differences in the exoproteome profile might serve as markers to discriminate the three groups of *S. aureus* isolates with different epidemiological backgrounds. Of note, some observed differences for accessory virulence genes that are apparently distinctive for the CA and HA isolates at the genome level, such as the genes for PVL and enterotoxins, were not reflected in the present proteomics analyses, because the respective proteins were not detected. This lack of detection may relate to their actual expression levels under the investigated conditions, or to the fact that the actual identification of proteins by MS depends on various factors, including the method of sample preparation, or the acquisition and analysis of the MS data. Yet, it should be noted that these proteins were identified in some other proteomic studies[40,41].

Knowledge of the sub-cellular localization of bacterial proteins can provide valuable insights into protein functions, especially in relation to colonization of the host, fitness and virulence [26]. In this respect, the attention is usually focused on actively secreted proteins that are synthesized with N-terminal signal peptides, because these include the major known virulence factors [26,27]. However, our proteomic analysis revealed only few significant differences in the detection of actively secreted proteins in the CA and HA groups that could be related to virulence (i.e. IsdA and IsdB), adhesion to host tissues (i.e. Ebps, Vwb), or immune evasion (i.e. SCIN). In addition, some other known virulence factors, such as CHIPS, the enterotoxin type D and the enterotoxin type A were uniquely identified in the CA^{DK}, HA^{DK}, and HA^{NLDE} isolates, respectively, irrespective of the growth phase.On the other hand, we observed major differences in the appearance of predicted cytoplasmic proteins in the exoproteomes of the investigated CA and HA isolates, where a higher number of cytoplasmic proteins was identified in the growth medium of the CA group than in the medium of the two HA groups during the exponential growth phase. Conversely, a higher number of cytoplasmic proteins was identified in the HA group than the CA group during the stationary growth phase. This difference can be interpreted in at least three ways. Firstly, there may be a difference in the timing of autolysis of cells [36,37] resulting in the early release of cytoplasmic proteins into the extracellular milieu by cells from the CA group. On the other hand, it is known that cell wall-associated and secreted proteases can degrade cytoplasmic proteins released into the

growth medium [36]. Hence a second possible explanation for the observed differences would be that the HA isolates are more proteolytic in the exponential growth phase than the CA isolates, leading to the observed differences due to degradation of liberated cytoplasmic proteins. In a third possible scenario, 'cytoplasmic' proteins are actively delivered into the growth medium via non-classical secretion mechanisms that are as yet ill-defined [33,34,37]. In this case, one would have to assume that the timing of non-classical secretion differs for CA and HA isolates. Inspection of the genome sequences of the investigated MRSA isolates showed that the genes for known autolysins (*atl*, *isaA*, *lytM*), proteases (*aur*, *sspA*, *sspB*, *sspC*, *spiA*, *spiB*, *spiC*, *spiD*, *spiE*, *spiF* and *IsaA*) and secretion pathways (*sec*, *tat* and type VII secretion) are intact with only few SNPs detectable in the coding and intergenic regions, as was the case for major gene regulators. None of the observed SNPs causes a premature stop of translation or a mutation that is known to be important for activity (data not shown). Of note, our comparative genome analysis suggests that the distinctive features of CA- and HA-MRSA isolates relate predominantly to the accessory genome, which might suggest a role of the respective genes at least in the timing of the extracellular appearance of predicted cytoplasmic proteins.

The principal component analysis (PCA) on the exoproteome data based on normalized spectral counts grouped the 15 investigated *S. aureus* isolates into two distinct groups. Herein, all the HA isolates formed a distinct cluster, whereas four of the six CA isolates formed a separate cluster. Importantly, the clustering of all HA^{DK} and $HA^{NL\text{-}DE}$ isolates in one group implies that our proteomics approach identifies a common signature of all investigated HA-MRSA isolates. Intriguingly, two isolates designated as CA grouped with the HA isolates, suggesting that these two "CA isolates" (D29, D61) may actually be hospital-adapted isolates that could have been propagated in the community. Consistent with this idea, our Northern blotting analysis showed that the transcript levels for exoproteins like FabF, RpoB and PSMα1-4 in the CA isolate D29 resembled more closely the respective profiles in the HA isolates. Yet, this was not the case for isolate D61, whose *fabF* and *rpoB* transcript levels matched with those of the CA isolates, while the *psmα1-4* transcript level was very low. An alternatively possibility is that the CA isolates D29 and D61 are genuine CA isolates, in which case our proteome analyses might highlight another distinguishing feature of the two clusters. Indeed, the distinction of CA and HA groups based on our PCA analysis seems to have predictive value for the growth behavior and survival of *S. aureus* in non-professional phagocytic epithelial cells. Clearly, the CA isolates displayed an increase in net cell number after internalization by epithelial cells compared with the HA isolates, whose bacterial count did not substantially increase following internalization. This finding would be fully in line with an earlier study that suggested better survival of CA-MRSA than of HA-MRSA inside human neutrophils [42]. Together, these findings imply that the cytoplasmic proteins identified in the exoproteome are indicative not only for the epidemiological behavior, but also could have an impact on the intracellular behavior of *S. aureus* within epithelial cells influencing their survival or replication capabilities. Of course, this does not exclude the possible involvement of known virulence factors with a clear role in virulence, such as phenol-soluble modulins, which were previously implicated in epidemiological behavior and intracellular survival [29,43–45], or the leukocidins PVL or LukAB/ED [46].

In recent years, increasing evidence has been obtained that particular cytoplasmic proteins may have different functions at intracellular and extracellular locations [47,48]. Such proteins are often regarded as 'moonlighting' proteins. Of note, the cytoplasmic proteins which we could consider here as potential moonlighting proteins are mostly proteins that are constitutively expressed at relatively high levels, and that have previously identified roles in processes such as sugar metabolism, adherence to host tissues, pathogenesis, and/or immune evasion [49–52]. Of note, a number of these potentially moonlighting proteins do not only occur in prokaryotes but also in eukaryotes, which could be suggestive of molecular mimicry where pathogens disguise themselves with a corona of factors that the human immune system does not recognize as non-self[50,53,54]. Altogether, the possible roles of moonlighting cytoplasmic proteins in the different epidemiology and intracellular survival of CA and HA USA300 isolates as highlighted in our present study would

be fully in line with recent studies where it was proposed that such proteins contribute to staphylococcal virulence[33,35,55,56].

We conclude that our present proteomics approach to identify exoproteome signatures, and the results obtained with this approach, open up novel avenues to study and predict the epidemiological behavior of clinical MRSA isolates. Clearly, our study is built on genetically closely related MRSA isolates with distinct epidemiological behavior. It will be an important challenge for future research to assess whether a similar distinction can be achieved when this approach is applied to genetically distantly related MRSA isolates with different epidemiological behavior in hospitals and the community.

Materials and Methods

Bacterial isolates

Relevant properties of the 15 MRSA isolates used for exoproteome analyses are listed in Supplementary Table 7. 12 isolates with the PFGE profile USA300 were collected by the Statens Serum Institut (Copenhagen, Denmark) in the period between 1999 and 2006 [21]. These 12 isolates included six CA-MRSA isolates with *spa* type t008 (referred to as CA^{DK}), and six HA-MRSA isolates with *spa* type t024 (referred to as HA^{DK}) [18,21]. The remaining three HA-MRSA isolates with *spa* types t008 or t024 (referred to as HA^{NL-DE} isolates) were collected in the period between 1996 to 2010 in hospitals located within the Dutch-German border region (EUREGIO)[19].

Whole genome sequencing of isolates and analysis

Whole genome sequencing of the investigated *S. aureus* isolates was performed on an Illumina MiSeq instrument and the Nextera® XT, 2x 250bp kit using the manufacturer`s standard protocols (Illumina, Inc, USA). DNA for the sequencing was extracted using the DNeasy Blood and Tissue kit (Qiagen, Valencia, CA, USA). The WGS datasets generated and/or analysed in the current study are available in the European Nucleotide Archive (ENA) repository under accession number ERP018940 (http://www.ebi.ac.uk/ena/data/view/PRJEB17079). Reads from the isolates were assembled using SPAdes [57], annotated using PROKKA [58], and the core and pan-genome of the isolates was estimated using ROARY [59]. The alignment of the core genome from ROARY was used as input to create a phylogenetic tree using RAxML [60] with 100 bootstrap supports. The phylogenetic tree and accessory genome heatmap were visualized using ggtree. Specifically, the phylogenetic tree was based on the variable positions in the core genome of the 15 isolates (2,334 genes). The accessory genome was defined as genes present in at most 70% of the isolates resulting in 992 accessory genes. Virulence genes were identified using VirulenceFinder [61]. The analysis of potential antimicrobial resistance genes was performed with the web-based ResFinder tool [62].

Bacterial cultivation for proteome sampling

Bacteria were grown overnight (14-16 h) at 37°C in 25 mL Trypton Soy Broth (TSB) under vigorous shaking (115 rpm). The cultures were then diluted into 25 mL pre-warmed RPMI 1640 medium supplemented with 2 mM glutamine (GE Healthcare/PAA, Little Chalfont, United Kingdom) to an OD_{600} of 0.05 and cultivation was continued under the same conditions. Exponentially growing cells with an OD_{600} of ~0.5 were re-diluted into 120 mL fresh, pre-warmed RPMI medium to a final OD_{600} of 0.05. The cultivation was continued until the cultures had reached 90 min within the stationary growth phase. Within this period, two time points were selected for sample collection. The first one was set at OD_{600} of ~0.5, corresponding to the exponential growth phase, and the second one was set at ~OD_{600} of 1.3, corresponding to approx. 90 min after entry into the stationary phase. At these two time points, 1.5 ml culture aliquots were collected. In brief, the collected aliquots were centrifuged for 10 min at 4°C and 8000×g with the subsequent application of a 0.22 µM filter step (GE Healthcare Systems, Little Chalfont, United Kingdom) to remove the remaining bacterial cells. The

extracellular proteins in the supernatant were precipitated with 10 % w/v TCA on ice at 4°C overnight. Finally, the precipitates were collected by centrifugation for 20 min at 4°C and 8000×g, washed with ice-cold acetone, and dried at room temperature. The dried protein pellets were stored at -20°C until further use. For each isolate two biological replicates were analyzed, which adds up to 24 exponential phase and 24 stationary phase samples for the CA^{DK} and HA^{DK} strains, respectively, and to 6 exponential phase and 6 stationary phase samples for the HA^{NL-DE} strains.

Sample preparation for proteome analysis

Dried protein samples were processed as described previously [54]. Briefly, protein pellets were dissolved in 50 mM ammonium bicarbonate buffer (Fluka, Buchs, Switzerland), reduced with 10 mM dithiothreitol (DuchefaBiochemie, Haarlem, the Netherlands) for 30 min, and alkylated with 10 mM iodoacetamide (Sigma-Aldrich, St. Louis, USA) for 30 min in the dark. To digest complex protein samples, 80 ng trypsin (Promega, Madison, USA) was added and the samples were incubated overnight at 37°C under static conditions. To stop the digestion, the samples were acidified with a final concentration of 0.1 % trifluoroacetic acid (TFA, Sigma-Aldrich, St. Louis, USA) and subsequently purified using ZipTips (Millipore, Billerica, USA). For this purpose, the tips were stepwise equilibrated with 30 µL acetonitrile (ACN, Fluka, Buchs, Switzerland), 30 µL 80 % ACN/0.1 % TFA, 50 % ACN/0.1 % TFA, 30 µL 30 % ACN/0.1 % TFA and finally 30 µL 0.1 % TFA. Peptides were bound to ZipTips by pipetting 10 times 10 µL of the sample. Impurities were removed by washing with 50 µL 0.1 % TFA and finally peptides were eluted with 20 µL 50 % ACN/0.1 % TFA and 20 µL 80 % ACN/0.1 % TFA. The final eluates were concentrated using a vacuum centrifuge (Eppendorf, Hamburg, Germany) and stored at 4°C until further use.

Mass spectrometry

Tryptic peptides were separated by reversed phase liquid chromatography (LC) coupled online to electrospray ionization mass spectrometry (ESI-MS) using an LTQ Orbitrap as described by Stobernack *et al*. [63]. Database searching was done with Sorcerer-SEQUEST 4 (Sage-N Research, Milpitas, USA). After extraction from the raw files, *.dta files were searched with Sequest against a target-decoy database with a set of common laboratory contaminants. The databases for the respective peptide/protein search were created from the genome sequences of the 15 investigated MRSA isolates. The RAST annotation file of these 15 MRSA isolates was used to create a non-redundant database comprising protein sequences of all isolates. Protein sequences that differed in only one amino acid were included in this database. Finally, validation of MS/MS-based peptide and protein identifications was performed with Scaffold v4.3.4 (Proteome Software, Portland, USA). Peptide identifications were accepted if they exceeded specific database search engine thresholds. SEQUEST identifications required at least deltaCn scores of greater than 0.1 and XCorr scores of greater than 2.2, 3.3 and 3.75 for doubly, triply and quadruply charged peptides, respectively. With these filter parameters, no false-positive hits were obtained, which was verified by a search against a concatenated target-pseudo reversed decoy database. Normalized spectral counts were obtained from the Scaffold file by considering a 99 % protein threshold, and a minimum of two peptides for each protein. The normalized spectral count data were exported from Scaffold and curated in Microsoft Excel before further analysis. The filtered MS data associated with this manuscript can be downloaded from the PRIDE partner repository of the ProteomeXchange Consortium using the following link: http://www.ebi.ac.uk/pride/archive/login (Username: reviewer84128@ebi.ac.uk, and Password: fzzQQiRF).

Prediction of protein localization, biological processes and molecular functions

Prediction of the subcellular localization of proteins that were identified by LC-MS/MS was performed using different bioinformatics tools. Since individual bioinformatics tools are not able to specifically predict all

possible localization sites of bacterial proteins [64], we used eight different computer programs, namely SignalP4.1 (http://www.cbs.dtu.dk/Services/SignalP/) [65], Phobius (http://www.ebi.ac.uk/Tools/pfa/phobius/)[66], Predisi (http://www.predisi.de/) [67], LipoP1.0 (http://www.cbs.dtu.dk/services/LipoP) [68], ProtComp9.0 (http://linux1.softberry.com/berry.phtml?topic=protcompan&group=programs&subgroup=proloc)[69], PSort 3.0.3b (http://www.psort.org/psortb/index.html) [70], TMHMM2.0c (http://www.cbs.dtu.dk/services/TMHMM) [71], and PSscan (http://prosite.expasy.org/scanprosite/) [72]. The settings used for of each program are specified in Supplementary Table 8. A detailed description of output parameters, scores and thresholds for each tag is presented in Supplementary Table 9. Voronoi treemaps to link quantitative proteomic data and functional classifications were created using the Paver software (DECODON GmbH, Greifswald, Germany) with the latest functional categorization of SEED database of *S. aureus* USA300_FPR3757 [73].

RNA isolation

Bacterial isolates were grown under the same condition as for the proteomics sample collection. 25 ml culture aliquots were collected for RNA isolation 90 min after entry into the stationary growth phase, corresponding to an OD_{600} of approx. 1.3. RNA was isolated from bacteria as described previously [74]. Briefly, ½ volume of frozen killing buffer (20 mM Tris/HCl [pH 7.5], 5 mM MgCl2, 20mM NaN3) was added to the bacterial culture, and bacterial cells were collected by centrifugation for 3 minutes, 8000 rpm at 4°C. The supernatant was discarded, and pellets were frozen in liquid nitrogen and stored at -80°C until further processing. Cell pellets were re-suspended in ice-cold killing buffer and transferred into Teflon vessels filled with liquid N_2 for disruption. Cells were then mechanically disrupted with a Mikro-Dismembrator S (Sartorius) for 2 min, 2600 rpm. The resulting powder was re-suspended in lysis solution that was pre-warmed at 50°C (4 M guanidine thiocyanate, 25 mM sodium acetate [pH 5.2], 0.5% N-laurylsarcosinate 40 [wt/vol]) by repeated up- and down-pipetting. Then, lysates were transferred into pre-cooled micro-centrifuge tubes, and frozen at -80°C.

Total RNA was isolated by phenol-chloroform extraction as described previously [74]. Samples were processed twice with an equal volume of acid phenol solution (Sigma-Aldrich, Zwijndrecht, the Netherlands), and mixed thoroughly on an Eppendorf tube shaker until completely thawed. The resulting suspension was then centrifuged for 5 min, 12000 rpm, and the supernatant was transferred into a fresh microcentrifuge tube. Next, samples were processed once with one volume of Chloroform/isoamyl alcohol, mixed well, and centrifuged for 5 min, 12000 rpm. RNA was precipitated from the supernatant by the addition of 1/10 volume of 3 M Na-Acetate, pH 5.2, and 0.8 ml of isopropanol. The precipitated RNA was washed once with 70 % RNase-free ethanol, and dissolved in RNase-free water.

Northern blot analysis

Northern blot analysis was performed as described previously [75]. Specific biotin-labelled RNA probes were generated by *in vitro* synthesis using a T7 RNA polymerase and Bio-16-UTP (Life Technologies). 3-10 µg of total RNA per lane was separated on 1.2 % denaturing agarose gels. Gene-specific transcripts were detected with the aid of biotin-labeled anti-sense RNA-probes. Fluorescent detection of the biotin- labelled probes was carried out using IRDye® 800CW Streptavidin (LI-COR Biosciences - GmbH) and the Odyssey Clx Imaging System (LI-COR Biosciences - GmbH) according to the instructions of the manufacturer. Primer sequences are listed in Supplementary Table 10.

Survival of bacteria upon epithelial cell infection

Cell lines and culture conditions

The human bronchial epithelial cell line 16HBE14o- was used to investigate the survival of MRSA isolates upon internalization. The epithelial cells were cultured in eukaryotic minimal essential medium (eMEM; 1x MEM without arginine and lysine; Costumer formulation, PromoCell GmbH, Heidelberg, Germany)

supplemented with 10 % (v/v) fetal calf serum (FCS; Biochrom AG, Berlin, Germany), 2 % (v/v) L-glutamine 200 mM (PAN-Biotech GmbH, Aidenbach, Germany) and 1 % (v/v) non-essential amino acids 100x (PAN-Biotech GmbH). The cells were seeded at a density of $1x10^5$ cells/cm^2 in CellStar® 12-well plates (Greiner Bio-One, Frickenhausen, Germany) and cultured for three days at 37°C, 5 % CO_2 in a humidified atmosphere after which they were ready for infection experiments.

Bacterial culture conditions

The bacteria were cultured in prokaryotic minimal essential medium (pMEM; 1x MEM without sodium bicarbonate; Invitrogen, Karlsruhe, Germany) supplemented with 1x non-essential amino acids (PAN-Biotech GmbH), 4 mM L-glutamine (PAN-Biotech GmbH), 10 mM HEPES (PAN-Biotech GmbH), 2 mM L-alanine, 2 mM L-leucine, 2 mM L-isoleucine, 2 mM L-valine, 2 mM L-aspartate, 2 mM L-glutamate, 2 mM L-serine, 2 mM L-threonine, 2 mM L-cysteine, 2 mM L-proline, 2 mM L-histidine, 2 mM L-phenyl alanine and 2 mM L-tryptophan (Sigma-Aldrich, Munich, Germany), adjusted to pH 7.4 and sterilized through filtration. Notably, for the overnight pre-culture 0.01 % of yeast extract was added.

Internalization procedure

The internalization of MRSA into epithelial cells was carried out as described previously by Pförtner *et al*. [76]. Briefly, bacterial cultures were inoculated from exponentially growing overnight cultures, starting at an inoculation OD_{600} of 0.05 and permitting growth until the mid-exponential phase at 37°C, 150 rpm in a shaking water bath. The bacterial numbers were determined by flow cytometry with a Guava easyCyte™ flow cytometer (MilliporePrior Billerica, MA, USA). Prior to infection, the numbers of epithelial cells were assessed by detaching them from the plates with trypsin-EDTA 0.25 % (Thermo Fisher Scientific, Waltham, USA), mixing with Trypan blue dye, and counting with a Countess® cell counter (Invitrogen, Karlsruhe, Germany). In order to infect epithelial cells with MRSA at a multiplicity of infection (MOI) of 1:25, the host cell medium was exchanged with the infection mix (MRSA diluted on eMEM, buffered with 2.9 µl sodium hydrogen carbonate [7.5 %] per ml of bacterial culture added) and incubated for one hour at 37°C, 5 % CO_2 in an incubator. Afterwards, the cell culture medium was exchanged with fresh eMEM containing 10 µg/ml lysostaphin, and this medium was exchanged every two days for long-term experiments.

Sampling of the 16HBE14o- cells was performed by detaching of the cells from the plate with trypsin-EDTA 0.25 %, and the collection of internalized bacteria was carried out through incubation of the plate with 0.05 % SDS for 5 min. Quantification of the intracellular MRSA isolates was performed by flow cytometry with a GUAVA®easyCyte (Merck Millipore, Darmstadt, Germany). To this end, the bacteria were stained with 0.2 µg/ml Vancomycin BODIPY FL (Thermo Fisher Scientific, Waltham, USA), and detected using a 488 nm laser for excitation as described [77]. The intracellular survival of each isolate was analyzed in independent duplicate experiments.

Graphical and statistical analyses

Volcano plot analyses were performed using GraphPad Prism version 6. Statistical analyses were performed using the Wilcoxon signed-rank test. A P-value of less than or equal to 0.05 was considered statistically significant. Principal component analysis (PCA) was performed using the Statistical Package for Social Science (SPSS) version 22. The component loading of the extracellular proteins from the 15 CA^{DK}, HA^{DK} and HA^{NL-DE} isolates was calculated both for growth medium fractions of cells in the exponential and stationary growth phases based on normalized spectral count. The Venn diagram was constructed using Venny version [78].

Ethics

The present research has no particular ethical implications.

Funding

This work was supported by the Graduate School of Medical Sciences of the University of Groningen [to S.A.M., L.M.P.M. and C.G], Deutsche Forschungsgemeinschaft Grant GRK1870 [to L.M.P. M. and U.V.], the People Programme (Marie Skłodowska-Curie Actions) of the European Union's Horizon 2020 Programme under REA grant agreement no. 642836 [to S.G., D.B. and J.M.v.D.], and Deutsche Forschungsgemeinschaft (SFB/TRR 34 framework) [to D.B. and A.O.].

Disclosure of Potential Conflicts of Interest

The authors declare that they have no financial and non-financial competing interests in relation to the documented research.

Acknowledgements

We thank Eliane Popa for critically reading the manuscript, and Alex Reder, Alex Friedrich and Girbe Buist for helpful discussions.

References

1. Kriegeskorte A, Peters G. Horizontal gene transfer boosts MRSA spreading. Nat Med. 2012; 18:662–3.
2. Wertheim HF, Melles DC, Vos MC, van Leeuwen W, van Belkum A, Verbrugh H a, Nouwen JL. The role of nasal carriage in *Staphylococcus aureus* infections. Lancet Infect Dis 2005; 5:751–62.
3. Lowy FD. Antimicrobial resistance: The example of *Staphylococcus aureus*. J. Clin. Invest.2003; 111:1265–73.
4. Chambers HF, Deleo FR. Waves of resistance: *Staphylococcus aureus* in the antibiotic era. Nat Rev Microbiol 2009; 7:629–41.
5. De Kraker ME, Wolkewitz M, Davey PG, Grundmann H. Clinical impact of antimicrobial resistance in European hospitals: Excess mortality and length of hospital stay related to methicillin-resistant *Staphylococcus aureus* bloodstream infections. Antimicrob Agents Chemother 2011; 55:1598–605.
6. De Kraker ME, Davey PG, Grundmann H. Mortality and hospital stay associated with resistant *Staphylococcus aureus* and *Escherichia coli* bacteremia: Estimating the burden of antibiotic resistance in Europe. PLoS Med 2011; 8.
7. Sun H, Wei C, Liu B, Jing H, Feng Q, Tong Y, Yang Y, Yang L, Zuo Q, Zhang Y, et al. Induction of systemic and mucosal immunity against methicillin-resistant *Staphylococcus aureus* infection by a novel nanoemulsion adjuvant vaccine. Int J Nanomedicine 2015; 10:7275–90.
8. Olaniyi R, Pozzi C, Grimaldi L, Bagnoli F. *Staphylococcus aureus*-associated skin and soft tissue infections: anatomical localization, epidemiology, therapy and potential prophylaxis. Curr Top Microbiol 2016; Epub ahead of print.
9. Missiakas D, Schneewind O. *Staphylococcus aureus* vaccines: Deviating from the carol. J Exp Med 2016; 213:1645–53.
10. Weber JT. Community-associated methicillin-resistant *Staphylococcus aureus*. Clin Infect Dis 2005; 41 Suppl 4:S269-72.
11. DeLeo FR, Otto M, Kreiswirth BN, Chambers HF. Community-associated meticillin-resistant *Staphylococcus aureus*. Lancet 2010; 375:1557–68.
12. Wang R, Braughton KR, Kretschmer D, Bach T-HL, Queck SY, Li M, Kennedy AD, Dorward DW, Klebanoff SJ, Peschel A, et al. Identification of novel cytolytic peptides as key virulence determinants for community-associated MRSA. Nat Med 2007; 13:1510–4.
13. David MZ, Daum RS. Community-associated methicillin-resistant *Staphylococcus aureus*: Epidemiology and clinical consequences of an emerging epidemic. Clin. Microbiol. Rev.2010; 23:616–87.
14. Lindsay JA, Holden MTG. *Staphylococcus aureus*: Superbug, super genome? Trends Microbiol.2004; 12:378–85.
15. Sabat AJ, Budimir A, Nashev D, Sá-Leão R, van Dijl J m, Laurent F, Grundmann H, Friedrich AW, ESCMID Study Group of Epidemiological Markers (ESGEM). Overview of molecular typing methods for outbreak detection and epidemiological surveillance. Euro Surveill 2013;18:20380.
16. Thurlow LR, Joshi GS, Clark JR, Spontak JS, Neely CJ, Maile R, Richardson AR. Functional modularity of the arginine catabolic mobile element contributes to the success of USA300 methicillin-resistant *Staphylococcus aureus*. Cell Host Microbe 2013; 13:100–7.
17. Diep BA, Gill SR, Chang RF, Phan TH, Chen JH, Davidson MG, Lin F, Lin J, Carleton HA, Mongodin EF, et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. Lancet 2006; 367:731–9.
18. Larsen AR, Goering R, Stegger M, Lindsay JA, Gould KA, Hinds J, Sørum M, Westh H, Boye K, Skov R. Two distinct clones of methicillin-resistant *Staphylococcus aureus* (MRSA) with the same USA300 pulsed-field gel electrophoresis profile: A potential pitfall for identification of USA300 community-

associated MRSA. J Clin Microbiol 2009; 47:3765–8.

19. Glasner C, Sabat AJ, Dreisbach A, Larsen AR, Friedrich AW, Skov RL, van Dijl JM. Rapid and high-resolution distinction of community-acquired and nosocomial *Staphylococcus aureus* isolates with identical pulsed-field gel electrophoresis patterns and *spa* types. Int J Med Microbiol 2013; 303:70–5.
20. Stam-Bolink EM, Mithoe D, Baas WH, Arends JP, Möller AVM. Spread of a methicillin-resistant *Staphylococcus aureus* ST80 strain in the community of the northern Netherlands. Eur J Clin Microbiol Infect Dis 2007; 26:723–7.
21. Larsen AR, Stegger M, Böcher S, Sørum M, Monnet DL, Skov RL. Emergence and characterization of community-associated methicillin- resistant *Staphyloccocus aureus* infections in Denmark, 1999 to 2006. J Clin Microbiol 2009; 47:73–8.
22. Otter JA, French GL. Community-associated meticillin-resistant *Staphylococcus aureus* strains as a cause of healthcare-associated infection. J. Hosp. Infect.2011; 79:189–93.
23. Francois P, Scherl A, Hochstrasser D, Schrenzel J. Proteomic approach to investigate MRSA. Methods Mol Biol 2007; 391:179–99.
24. Mäder U, Nicolas P, Depke M, Pané-Farré J, Debarbouille M, van der Kooi-Pol MM, Guérin C, Dérozier S, Hiron A, Jarmer H, et al. *Staphylococcus aureus* Transcriptome Architecture: From Laboratory to Infection-Mimicking Conditions. PLOS Genet 2016; 12:e1005962.
25. Muers M. Gene expression: Transcriptome to proteome and back to genome. Nat Rev Genet 2011; 12:518.
26. Sibbald MJ, Ziebandt AK, Engelmann S, Hecker M, de Jong A, Harmsen HJM, Raangs GC, Stokroos I, Arends JP, Dubois JYF, et al. Mapping the pathways to staphylococcal pathogenesis by comparative secretomics. Microbiol Mol Biol Rev 2006; 70:755–88.
27. Ziebandt AK, Kusch H, Degner M, Jaglitz S, Sibbald MJJB, Arends JP, Chlebowicz MA, Albrecht D, Pantuček R, Doškar J, et al. Proteomics uncovers extreme heterogeneity in the *Staphylococcus aureus* exoproteome due to genomic plasticity and variant gene regulation. Proteomics 2010; 10:1634–44.
28. Koymans KJ, Vrieling M, Gorham RD, van Strijp JAG. Staphylococcal Immune Evasion Proteins: Structure, Function, and Host Adaptation. Curr Top Microbiol Immunol 2016; Epub ahead of print.
29. Spaan AN, Surewaard BG, Nijland R, van Strijp JA. Neutrophils versus *Staphylococcus aureus*: a biological tug of war. Annu Rev Microbiol 2013; 67:629–50.
30. He QY, Chiu JF. Proteomics in biomarker discovery and drug development. J Cell Biochem 2003; 89:868–86.
31. Tenover FC, Goering R V. Methicillin-resistant *Staphylococcus aureus* strain USA300: Origin and epidemiology. J. Antimicrob. Chemother.2009; 64:441–6.
32. Antelmann H, Tjalsma H, Voigt B, Ohlmeier S, Bron S, Van Dijl JM, Hecker M. A proteomic view on genome-based signal peptide predictions. Genome Res 2001; 11:1484–502.
33. Ebner P, Rinker J, Götz F. Excretion of cytoplasmic proteins in *Staphylococcus* is most likely not due to cell lysis. Curr Genet 2016; 62:19–23.
34. Wang G, Xia Y, Song X, Ai L. Common Non-classically Secreted Bacterial Proteins with Experimental Evidence. Curr. Microbiol.2016; 72:102–11.
35. Götz F, Yu W, Dube L, Prax M, Ebner P. Excretion of cytosolic proteins (ECP) in bacteria. Int. J. Med. Microbiol. 2015; 305:230–7.
36. Krishnappa L, Dreisbach A, Otto A, Goosens VJ, Cranenburgh RM, Harwood CR, Becher D, Van Dijl JM. Extracytoplasmic proteases determining the cleavage and release of secreted proteins, lipoproteins, and membrane proteins in *Bacillus subtilis*. J Proteome Res 2013; 12:4101–10.
37. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. BMC Microbiol 2005; 5:58.

38. Liew YK, Hamat RA, Belkum A Van, Chong PP, Neela V. Comparative exoproteomics and host inflammatory response in *Staphylococcus aureus* skin and soft tissue infections, bacteremia, and subclinical colonization. Clin Vaccine Immunol 2015; 22:593–603.
39. Liew YK, Hamat RA, Nordin SA, Chong PP, Neela V. The exoproteomes of clonally related *Staphylococcus aureus* strains are diverse. Ann Microbiol 2015; 65:1809–1813.
40. Cassat JE, Hammer ND, Campbell JP, Benson MA, Perrien DS, Mrak LN, Smeltzer MS, Torres VJ, Skaar EP. A secreted bacterial protease tailors the *Staphylococcus aureus* virulence repertoire to modulate bone remodeling during osteomyelitis. Cell Host Microbe 2013; 13:759–72.
41. Monteiro R, Hébraud M, Chafsey I, Chambon C, Viala D, Torres C, Poeta P, Igrejas G. Surfaceome and exoproteome of a clinical sequence type 398 methicillin resistant *Staphylococcus aureus* strain. Biochem Biophys Reports 2015; 3:7–13.
42. Voyich JM, Braughton KR, Sturdevant DE, Whitney AR, Saïd-Salim B, Porcella SF, Long RD, Dorward DW, Gardner DJ, Kreiswirth BN, et al. Insights into mechanisms used by *Staphylococcus aureus* to avoid destruction by human neutrophils. J Immunol 2005; 175:3907-19.
43. Geiger T, Francois P, Liebeke M, Fraunholz M, Goerke C, Krismer B, Schrenzel J, Lalk M, Wolz C. The stringent response of *Staphylococcus aureus* and its impact on survival after phagocytosis through the induction of intracellular PSMs expression. PLoS Pathog 2012; 8.
44. Cheung GYC, Joo HS, Chatterjee SS, Otto M. Phenol-soluble modulins - critical determinants of staphylococcal virulence. FEMS Microbiol. Rev.2014; 38:698–719.
45. Surewaard BGJ, De Haas CJC, Vervoort F, Rigby KM, Deleo FR, Otto M, Van Strijp JAG, Nijland R. Staphylococcal alpha-phenol soluble modulins contribute to neutrophil lysis after phagocytosis. Cell Microbiol 2013; 15:1427–37.
46. Dumont AL, Yoong P, Surewaard BGJ, Benson MA, Nijland R, van Strijp JAG, Torres VJ. *Staphylococcus aureus* elaborates the leukotoxin LukAB to mediate escape from within human neutrophils. Infect Immun 2013; 81:1830–41.
47. Radisky DC, Stallings-Mann M, Hirai Y, Bissell MJ. Single proteins might have dual but related functions in intracellular and extracellular microenvironments. Nat Rev Mol Cell Biol 2009; 10:228–34.
48. Huberts DH, van der Klei IJ. Moonlighting proteins: An intriguing mode of multitasking. Biochim Biophys Acta - Mol Cell Res 2010; 1803:520–5.
49. Bonar E, Wójcik I, Wladyka B. Proteomics in studies of *Staphylococcus aureus* virulence. Acta Biochim. Pol.2015; 62:367–81.
50. Kainulainen V, Korhonen TK. Dancing to another tune-adhesive moonlighting proteins in bacteria. Biology 2014; 3:178–204.
51. Wang G, Xia Y, Cui J, Gu Z, Song Y, Chen YQ, Chen H, Zhang H, Chen W. The roles of moonlighting proteins in bacteria. Curr. Issues Mol. Biol.2014; 16:15–22.
52. Otto A, van Dijl JM, Hecker M, Becher D. The *Staphylococcus aureus* proteome. Int. J. Med. Microbiol.2014; 304:110–20.
53. Wang W, Jeffery CJ. An analysis of surface proteomics results reveals novel candidates for intracellular/surface moonlighting proteins in bacteria. Mol Biosyst 2016; 12:1420–31.
54. Dreisbach A, Hempel K, Buist G, Hecker M, Becher D, Van Dijl JM. Profiling the surfacome of *Staphylococcus aureus*. Proteomics 2010; 10:3082–96.
55. Henderson B, Martin A. Bacterial moonlighting proteins and bacterial virulence. Curr Top Microbiol Immunol 2013; 358:155–213.
56. Ebner P, Rinker J, Nguyen MT, Popella P, Nega M, Luqman A, Schittek B, Di Marco M, Stevanovic S, Götza F. Excreted cytoplasmic proteins contribute to pathogenicity in *Staphylococcus aureus*. Infect Immun 2016; 84:1672–81.

57. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 2012; 19:455–77.
58. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 2014; 30:2068–9.
59. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015; 31:3691–3.
60. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014; 30:1312–3.
61. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol 2014; 52:1501–10.
62. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 2012; 67:2640–4.
63. Stobernack T, Glasner C, Junker S, Gabarrini G, de Smit M, de Jong A, Otto A, Becher D, van Winkelhoff AJ, van Dijl JM. The Extracellular Proteome and Citrullinome of the Oral Pathogen *Porphyromonas gingivalis*. J Proteome Res 2016; 15:4532-4543.
64. Gardy JL, Brinkman FSL. Methods for predicting bacterial protein subcellular localization. Nat Rev Microbiol 2006; 4:741–51.
65. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 2011; 8:785–6.
66. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004; 338:1027–36.
67. Hiller K, Grote A, Scheer M, Münch R, Jahn D. PrediSi: Prediction of signal peptides and their cleavage positions. Nucleic Acids Res 2004; 32.
68. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 2003; 12:1652–62.
69. Clark HF, Gurney AL, Abaya E, Baker K, Baldwin D, Brush J, Chen J, Chow B, Chui C, Crowley C, et al. The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: A bioinformatics assessment. Genome Res 2003; 13:2265–70.
70. Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Cenk Sahinalp S, Ester M, Foster LJ, et al. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 2010; 26:1608–15.
71. Kahsay RY, Gao G, Liao L. An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. Bioinformatics 2005; 21:1853–8.
72. de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. ScanProsite: Detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 2006; 34.
73. Bernhardt J, Michalik S, Wollscheid B, Völker U, Schmidt F. Proteomics approaches for the analysis of enriched microbial subpopulations and visualization of complex functional information. Curr. Opin. Biotechnol.2013; 24:112–9.
74. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in *Bacillus subtilis*. Science 2012; 335:1103–6.
75. Homuth G, Masuda S, Mogk A, Kobayashi Y, Schumann W. The dnaK operon of *Bacillus subtilis* is heptacistronic. J Bacteriol 1997; 179:1153–64.
76. Pförtner H, Wagner J, Surmann K, Hildebrandt P, Ernst S, Bernhardt J, Schurmann C, Gutjahr M, Depke

M, Jehmlich U, et al. A proteomics workflow for quantitative and time-resolved analysis of adaptation reactions of internalized bacteria. Methods 2013; 61:244–50.

77. Hildebrandt P, Surmann K, Salazar MG, Normann N, Völker U, Schmidt F. Alternative fluorescent labeling strategies for characterizing gram-positive pathogenic bacteria: Flow cytometry supported counting, sorting, and proteome analysis of *Staphylococcus aureus* retrieved from infected host cells. Cytom Part A 2016;89:932-940.
78. Oliveros JC. VENNY. An interactive tool for comparing lists with Venn Diagrams. [Internet]. BioinfoGP of CNB-CSIC2007; :http://bioinfogp.cnnb.csic.es/tools/venny/index.ht. Available from: http://bioinfogp.cnnb.csic.es/tools/venny/index.html.

CHAPTER 5

EXPLANATION AND PREDICTION OF SIGNAL PEPTIDE EFFICIENCY: A MACHINE LEARNING MODEL TRAINED ON HIGH-THROUGHPUT DATA

Stefano Grasso†, Valentina Dabene†, Margriet M.W.B. Hendriks, Priscilla Zwartjens, René Pellaux, Martin Held, Sven Panke, Jan Maarten van Dijl#, Andreas Meyer#, Tjeerd van Rij#

†,# These authors contributed equally

Submitted for publication in *Nature Communications*

Abstract

Secreted proteins find important applications in various fields of the bioeconomy, from healthcare, to food and chemical industries[1]. To direct secretory proteins towards the extracellular space, they are synthesized with an N-terminal signal peptide (SP)[2,3]. Despite the universal importance in all kingdoms of life, many years of study, and industrial relevance, it is still not known what makes certain SPs more effective than others. To elucidate relevant features of the SP sequence that influence secretion efficiency, we screened a library of ~12,000 rationally designed SPs through a novel miniaturized high-throughput (HT) assay, combined the results with a simple machine learning model, and explained the model. By both quantifying the impact of SP features on secretion efficiency, and predicting the efficiency of designed and pseudo-random SPs, we devised a Design-Build-Test-Learn (DBTL) cycle to define critical SP features. Our study presents the blueprint for a structured approach that allows in silico design of optimal SPs.

Supplementary files avaialble at: https://github.com/grassoste/Thesis-supplementary-files

Main

The general protein secretion (Sec) machinery and associated SPs are responsible for translocation of the majority of bacterial proteins across the cytoplasmic membrane[2,4,5]. Because of its high capacity, the Sec pathway of various organisms was engineered to generate microbial cell factories for production of secreted proteinaceous products[6,7]. Monoderm Gram-positive bacteria, like *Bacillus subtilis*, are preferred for this purpose, as proteins only need to pass a single membrane, which eases the secretion process and subsequent recovery of bulk amounts of protein from the fermentation broth[2,8,9]. To target a protein of interest (POI) to the extracellular space, it is common practice to fuse a SP to its N-terminus. The SP structure has been elucidated, showing that it is universally composed of a positively charged N-region, a hydrophobic *α*-helical H-region, and a C-region. The latter encompasses an Ala-X-Ala motif (with X being any amino acid) for cleavage by a signal peptidase during or after translocation[10–12]. Today, the presence of a SP within a protein sequence can be reliably predicted[13]. However, there are no tools to foresee the secretion efficiency of a given SP-protein combination[3]. Thus, finding the best SP to secrete a POI is based on trial-and-error. Previous studies tested a limited number of natural SP variants (i.e. up to 10^2) and analyzed the relationships between secretion efficiency and a few SP features[14–16]. Such studies showed that the SP-POI match plays a crucial role in determining secretion efficiency[12], but did not unveil which particular SP features are the most significant.

Here we aimed to elucidate the relevant physicochemical features that determine secretion efficiency of a specific POI, namely the *α*-amylase AmyQ from *Bacillus amyloliquefaciens*, within a defined setting (i.e. specific growth and assay conditions), taking into account not only the amino acid (AA) but also the nucleotide sequences. To achieve this aim, we devised a workflow, based on the DBTL cycle approach, which uses HT quantification of the secretion efficiency to generate the training dataset for a machine learning (ML) model. Since ML models are black-boxes[17], our model was coupled with a post hoc model specific explanation of it, by means of TreeSHAP[18–21] (hereafter SHAP). This allows a dissection of the impact of individual physicochemical features describing the SPs over the modelled secretion efficiency, delivering an explainable and understandable model[22]. In doing so, we assayed the largest design space for SPs to date. Specifically, the sampling space and the variance covered by the naturally occurring SP sequences was rationally expanded with regard to the potentially relevant physicochemical features.

Starting from a selection of 134 known wild-type SPs from *B. subtilis*, a library of 11,643 unique SPs (hereafter SP-library; Supplementary Table 1) was designed. We applied a rational approach, individually modifying 7 specific features on 94 designated levels (Supplementary Table 2), while concomitantly minimizing their influence over related ones (e.g. editing the charge while avoiding significant change in hydrophobicity). In the present study, each SP was treated both as a single sequence and as 4 juxtaposed segments (i.e. the 3 classical aforementioned regions plus a short stretch of 3 AA after the cleavage site). As outlined in Figure 1, the designed SP-library was introduced into *B. subtilis* strain DB104 using a genome-integrating vector and a total of 160,000 clones was harvested to achieve a 10X coverage of the SP-library. HT screening of all generated clones was effectively performed by their compartmentalization in nanoliter reactors (NLRs)[23,24]. Inspired by the principle of the starch hydrolysis test[25], the secretion efficiency associated with each SP variant was determined by measuring the amylolytic degradation of fluorescein-labelled starch, co-encapsulated with the cells in NLRs (Figure 1a and 1b). Green fluorescence of each NLR was assessed by flow cytometry, with secretion of AmyQ being reflected in decreased signal intensity (Supplementary Figure 1 and Supplementary Figure 2). Additionally, to take into account the size of each growing colony, NLRs were incubated with Nile red, a hydrophobic red dye that interacts with the cell membrane[26] and enables fluorescent biomass determination. Occupied NLRs were classified in 10 equally populated bins, based on the relative secretion efficiency, defined as enzymatic activity per biomass unit (i.e. the ratio between green and red signals). Event collection was followed by DNA sequencing to determine the abundance of each SP

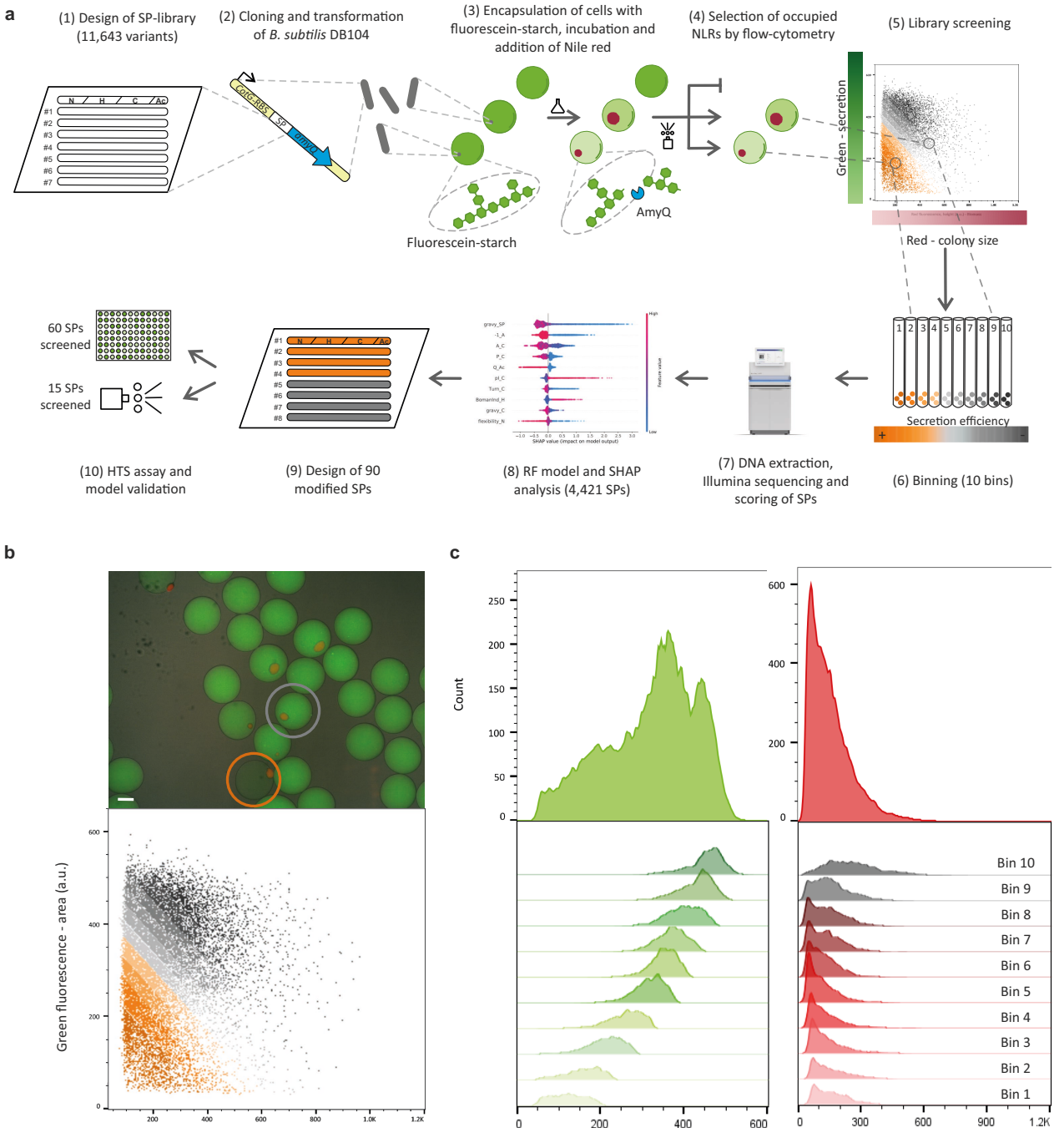

a
(1) Design of SP-library (11,643 variants)
(2) Cloning and transformation of B. subtilis DB104
(3) Encapsulation of cells with fluorescein-starch, incubation and addition of Nile red
(4) Selection of occupied NLRs by flow-cytometry
(5) Library screening
Green - secretion
Red - colony size
Fluorescein-starch
AmyQ
Secretion efficiency
(6) Binning (10 bins)
(7) DNA extraction, Illumina sequencing and scoring of SPs
(8) RF model and SHAP analysis (4,421 SPs)
(9) Design of 90 modified SPs
(10) HTS assay and model validation
60 SPs screened
15 SPs screened
b
Green fluorescence - area (a.u.)
c
Count
Bin 10
Bin 9
Bin 8
Bin 7
Bin 6
Bin 5
Bin 4
Bin 3
Bin 2
Bin 1

Figure 1. High throughput characterization of the signal peptide (SP) library. (a) Experimental workflow: (1) A library of approximately 12,000 signal peptide (SP) variants was designed by modifying key features (e.g. charge, length, hydrophobicity); (2) the corresponding pool of oligonucleotides was cloned in frame with the sequence coding for mature AmyQ, and integrated into the amyE locus of *B. subtilis* DB104. (3) The clones were embedded in hydrogel beads, referred to as nanoliter reactors (NLRs), containing fluorescein-starch (mean diameter of 500 µm; average occupation of 0.3 bacterial cells per NLR). During incubation in culture medium, single cells grew into microcolonies and secreted AmyQ, which degrades the fluorescein-starch into (still fluorescent) low molecular weight fragments that are lost from the NLR by diffusion. After incubation, biomass in the NLRs was labelled by adding Nile Red, a membrane-specific red fluorescent dye, and the NLRs were evaluated in 2 steps using a large particle flow cytometer. (4) Firstly, all empty NLRs were identified and discarded; (5-6) secondly, occupied NLRs were sequentially sorted into 10 bins, based on their green to red signal ratio. The green fluorescence signal is inversely proportional to the amount of secreted amylase (AmyQ) in the NLR; the red signal is instead directly proportional to the colony size. Therefore, clones with a high secretion efficiency are located in the lower left corner of the dot plot (5) and have a low bin number. (7) DNA from the NLRs of each bin was recovered and SP occurrence in any given bin was determined by NGS, leading to the construction of a frequency table of SPs across bins, used to calculate the secretion efficiency of each SP variant as a weighted average (WA). (8) WA values were subsequently combined with the features describing each SP to train a Random Forest (RF) regressor model. The RF model was then studied using SHAP for explanation and quantification of the impact of each feature on the model output (i.e. WA). (9) Information obtained by combining the RF model with the SHAP analysis was used to generate new SP variants with defined secretion levels to validate the model. (10) The designed validation sequences were processed following the same high-throughput (HT) screen, yet individually and not as a library. The secretion of amylase was quantified both with a microtiter plate assay (60 SPs) and by the NLR-based screening protocol (15 SPs), and the results were compared to the predictions. **(b)** Top: Overlay of bright-field and fluorescence microscopy images of NLRs after incubation in medium. Empty NLRs (no red dot) show a homogenous green fluorescence profile (no starch degradation), while NLRs harboring a colony (red dot) show different degrees of fluorescein-labelled starch degradation (orange circle: high secretor; grey circle: low secretor). Scale bar: 200 µm. Bottom: Dot plot representing all occupied NLRs from one experiment (approximately 20,000 NLRs). The gating applied during the second sorting step is depicted in orange-grey colour codes, which defines bins with distinct AmyQ secretion levels. Classification in bins based on the green to red ratio was set to deliver bins with equally sized NLR populations, i.e. 10% of library clones in each group. All events falling in one bin were sorted in bulk. **(c)** Green and red fluorescence profiles of all sorted events from one experiment (approximately 20,000 NLRs), both as a whole population (i.e. occupied NLRs; top panel) and divided into the 10 equally-sized bins (lower panel).

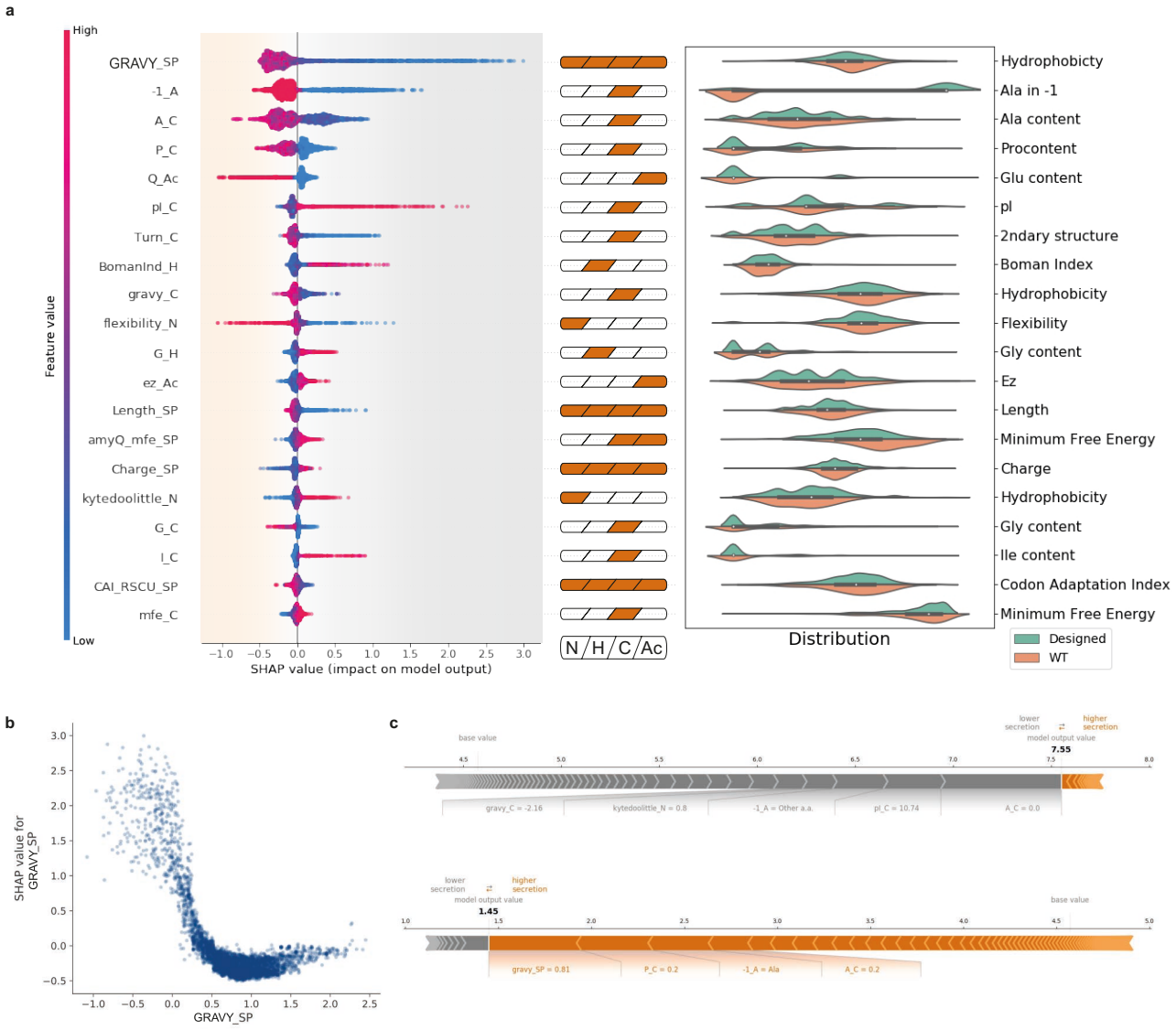

a
High
Low
Feature value
GRAVY_SP
-1_A
A_C
P_C
Q_Ac
pI_C
Turn_C
BomanInd_H
gravy_C
flexibility_N
G_H
ez_Ac
Length_SP
amyQ_mfe_SP
Charge_SP
kytedoolittle_N
G_C
I_C
CAI_RSCU_SP
mfe_C
−1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
SHAP value (impact on model output)
N/H/C/Ac
Hydrophobicty
Ala in -1
Ala content
Procontent
Glu content
pI
2ndary structure
Boman Index
Hydrophobicity
Flexibility
Gly content
Ez
Length
Minimum Free Energy
Charge
Hydrophobicity
Gly content
Ile content
Codon Adaptation Index
Minimum Free Energy
Distribution
Designed
WT
b
SHAP value for GRAVY_SP
GRAVY_SP
c
base value
lower secretion
higher secretion
model output value
7.55
gravy_C = -2.16
kytedoolittle_N = 0.8
-1_A = Other a.a.
pI_C = 10.74
A_C = 0.0
1.45
gravy_SP = 0.81
P_C = 0.2
-1_A = Ala
A_C = 0.2

Figure 2. Summary of the model description. (a) Left panel. SHAP summary plot of the 20 most impactful features, where a high dispersion of SHAP values on the abscissa indicates a broad effect of the respective feature on the model. Each data point represents a specific SP, the color of the data point indicates the value of that feature in the feature-specific scale, and the position on the abscissa indicates the SHAP value for that particular feature. SHAP values for the whole dataset sum up to the base value of the model (4.45 WA), which represents the average model output calculated over the 4,421 selected SPs. Thus, the sum of SHAP values for each individual feature of a specific SP will result in the output predicted by the RF model. Positive SHAP values indicate a negative impact on the model outcome, and vice versa. For example, 'GRAVY_SP' has the strongest impact on the model with low values (in blue) having a negative impact, and intermediate to high values (in red) having a positive impact on secretion. Middle panel: cartoon highlighting which part of the SP a particular feature refers to. Notably, the most impactful features belong either to regions close to the signal peptidase cleavage site (the C- and Ac- regions) or cover the entire SP. Right panel: Distributions of the top-20 features, in their feature-specific scales, for both the 134 wild-type SPs of *B. subtilis* 168 (WT; in orange) and the 4,421 valid SPs (in green). A brief description of each feature is presented as a label on the right (for full descriptions see Supplementary Table 2). **(b)** SHAP-dependence plot for 'GRAVY_SP', which is a bidimensional representation of the information summarized by the first line of Figure 2a. On the abscissa GRAVY index is represented, with negative values indicating low hydrophobicity, and positive values high hydrophobicity; on the ordinate SHAP values are displayed, with negative values indicating a beneficial effect on protein secretion, and positive values indicating a detrimental effect. Vertical dispersion of SHAP values for similar GRAVY indexes can be explained through the interaction effect between features (second order interactions are captured by SHAP interaction values, summarized in Supplementary Figure 8 and Supplementary Figure 9). To exemplify, the high variability visible in the negative range of the GRAVY index is to be attributed mainly to the feature '-1_A' (Supplementary Figure 7). From this graph it is possible to conclude that a very low hydrophobicity is predicted to have a strong negative impact on protein secretion, while a GRAVY index value of around 1.0 is predicted to show the most favorable impact. **(c)** SHAP force plot of a poorly (Top panel) and a good SP variant (Bottom panel) showing the values of the most relevant features with their relative impact on secretion efficiency. Each segment is sized proportionally to its impact on the model, their summation is equal to the difference between the base value (4.45 WA for all SPs) and the output value (in this case WAs of 7.55 and 1.45, for respectively the top and bottom panels). Features colored in gray have a negative impact on the secretion efficiency of the specified SP, while features colored in orange have a positive impact. For example, 'GRAVY_SP', with a value 0.81, has a positive impact.

variant in each bin, and, as a control, in the whole library at two different steps of the workflow (i.e. after transformation, and before sorting the NLRs into bins). Occurrence values were used to generate a weighted average (WA), assuming equidistance between bins, and this WA was ultimately used as a score for each SP.

To confirm the reliability of the data acquired through the NLR-based enzymatic assay (see Supplementary Figure 2 for initial validation), we compared the HT screening[27] methodology to two available assays: a microtiter plate (MTP) format using a synthetic substrate and the starch hydrolysis test using agar plates. We randomly picked 95 colonies producing fusions of the 4,421 selected SPs to AmyQ and measured the relative activity compared to AmyQ with its native SP using the MTP assay. The collected data showed a correlation with r=-0.50 between the NLR- and MTP-based activity assays (Figure 3a). A significant fraction (i.e. 73) of the selected 95 variants could not be measured using the standard MTP format and, even after an attempt to optimize its sensitivity, only 14 additional variants could be characterized (Supplementary Table 3). Therefore, we applied also a classical starch hydrolysis test on agar plates to verify the low-secreting variants, which validated the superior sensitivity of our NLR-based enzymatic assay (Supplementary Figure 3).

As shown by sequencing, 92% of the 11,643 unique rationally designed SPs were successfully introduced into *B. subtilis*, while 83% was retrieved after screening (Supplementary Table 4). Such reduction may relate to SP-dependent impaired growth and the resulting high background-to-noise ratios for small colonies. In addition, a strict threshold number for reliable reads and perfect mapping was defined, which reduced the number of SPs to be used to train and test our ML model to 4,421.

Next, we addressed and reduced the number of physicochemical features. Starting from an initial set of 267 features, 156 informative features were retained to describe each SP (Supplementary Table 2). This step removed features either presenting no variability or exhibiting a high correlation with another feature in the training dataset. A further reduction of dimensionality proved to be unnecessary and unwanted, as the PCA analysis showed that each of the principal components contributed to the explained variation (Supplementary Figure 4). Additionally, the same number of components was necessary to describe the whole variance of the 11,643 unique rationally designed SPs and the 4,421 informative SPs, indicating that, despite the loss in the total number of data points, there was no loss in the variation of the dataset. In contrast, the PCA showed that, to explain the same variation, more principal components are needed within the designed library than for the wild-type set of SPs, underpinning the improvement of the assayed space gained with our design (Supplementary Figure 4). The array of 156 features is thus to be considered as the independent variable, and the single value of secretion efficiency, the WA, as the dependent one.

Three quarters of our dataset were used to train a random forest (RF) regression algorithm, resulting in a mean squared error (MSE) of 1.75 WA, while the remaining quarter was used as a test set, resulting in an MSE of 1.22 WA (Supplementary Figure 5). After this first validation, we proceeded to provide explanations for the RF model predictions. Due to the complexity in explaining such a developed RF model, SHAP[18–20] was used to extract information about the importance of the features, and their interaction effects (Supplementary Figure 6). Here, one should bear in mind that the presented model, as any model, despite its predictive power, is conditional on the provided input and does not necessarily provide causal information about features.

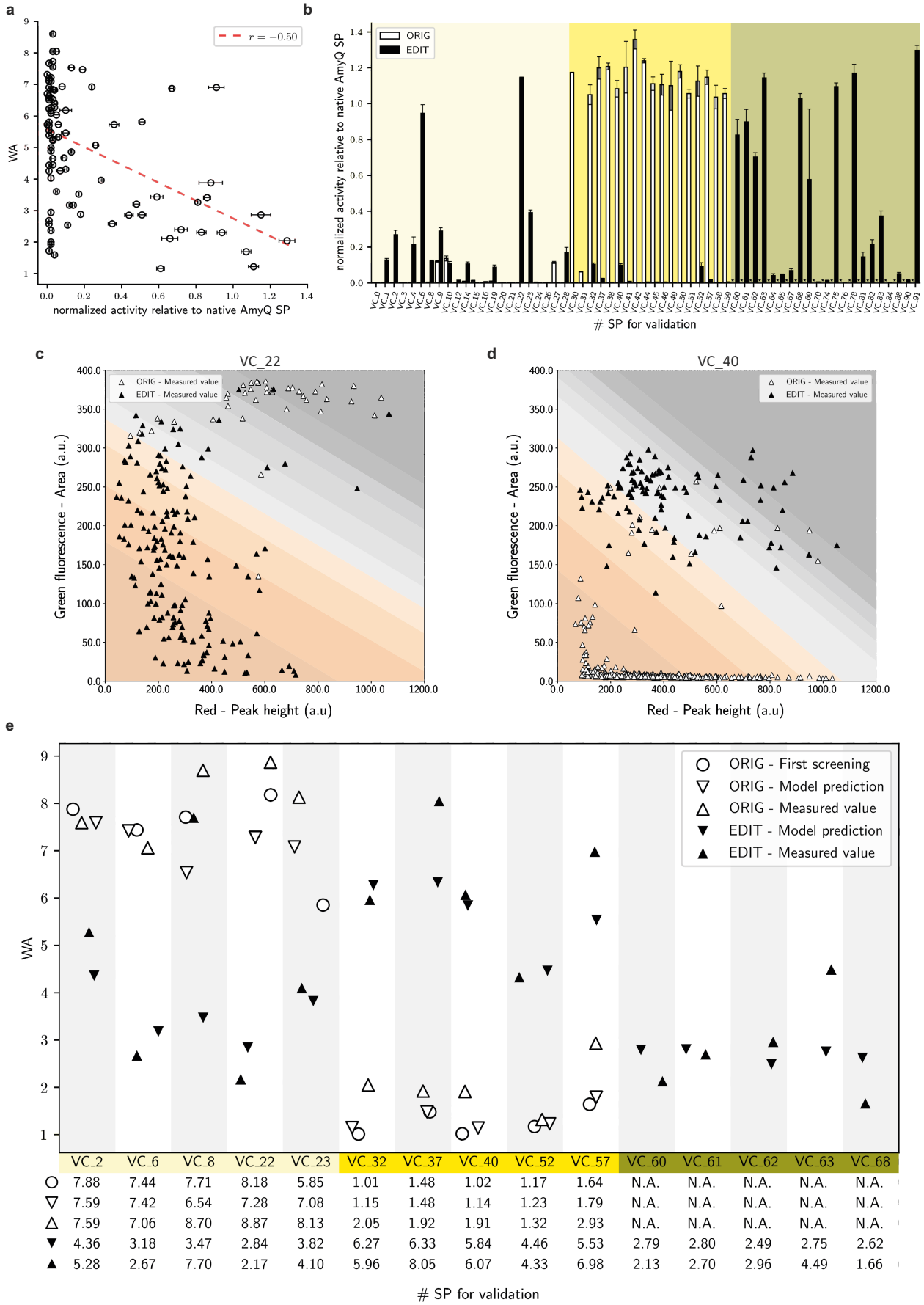
Due to the large number of features fed into the model and the notable amount of information provided by SHAP, only the most relevant and representative findings are highlighted and discussed. To fully explore the model, a Jupyter notebook (https://github.com/grassoste/SP-secretion-efficiency-Jupyter-notebook) and an interactive tool (File S1) are provided. The 20 most impactful features in our model are shown in Figure 2a. Some of these features were already documented in literature[3,28,29], for instance the overall hydrophobicity of the SP ('GRAVY_SP'), the presence of a helix breaking residue at the end of the H-region ('P_C' and 'Turn_C'), or the similarity of the cleavage site sequence to its consensus sequence (e.g. '-1_A' and 'A_C'). However, even for such known features, the impact on secretion could so far only be qualitatively estimated based on their distributions in wild-type SPs. With our approach, however, a more precise quantification is now

achieved, establishing favorable, neutral or detrimental values and their impact on the predicted secretion efficiency. Additionally, it is possible to determine, within the model, what kind of relationship particular features have with secretion efficiency (e.g. linear, sigmoidal, monotonic). This can be clearly illustrated with the dependency plot for a simple feature, such as 'GRAVY_SP' (Figure 2b), whose wild-type distribution is known and only includes positive values[29,30]. Our model analysis underlines the notion that functional positive GRAVY values are favorable, while a negative GRAVY value will be detrimental. Because not present in wild-type SPs, negative GRAVY values had thus far not been tested. Likewise, the model describes for the first time a quantitative estimate of the impact of these values on secretion efficiency (Figure 2b). Moreover, thanks to the applied segmentation approach (i.e. 4 juxtaposed regions for each SP), it is now possible to observe how some features have different relevance, depending on whether we consider the whole SP or only a single region. This is clearly exemplified by the feature 'Charge_SP', for which an overall value lower than +2 increases secretion efficiency, and a slightly negative charge is even more favorable. On the contrary, inspection of the charge of the N-region, represented by 'Flexibilty_N' (Supplementary Table 2), shows that values close to +2, or higher, favor secretion. Analogously, different features describing the same region can influence each other's impact. For instance, in the H-region, the feature 'BomanInd_H', which positively correlates with charge and negatively with hydrophobicity (Supplementary Table 1 and Supplementary Table 2), shows that a high level of hydrophobicity (low 'BomanInd_H') in this region can favor secretion. At the same time, judged by the feature 'G_H' (Gly content in the H-region), it appears that a high level of hydrophobicity related to a high Gly content is not favorable, most likely because Gly reduces the stability of α-helices. Notably, because of the applied feature selection process, one feature (e.g. 'BomanInd_H') may be representative of similar features (e.g. 'pI_H' and 'Charge_H'), which sets a limit to our immediate understanding of the influence of some properties. Nonetheless, with the present approach, we are able to retain, explain and trace back to their correlating counterparts, physicochemical properties of SPs, rather than less biologically significant indicators (e.g. principal components of a PCA, or the D-score from SignalP[14,31]).

With SHAP it is possible to analyze and quantify pairwise interactions between features, which explains why equal values of the same feature can influence the model to different extents. For instance, the vertical dispersion of 'GRAVY_SP', which is especially pronounced for negative hydrophobicity values (Figure 2B), is to be attributed to the feature '-1_A' (Supplementary Figure 7). Furthermore, overall interactions seem to play a minor role within our model, as the most impactful interaction ('Q_Ac'-'A_C') has limited impact on the overall output (Figures S8 and S9). One possible explanation is that interactions occur at orders higher than the second, and as such would not be represented in the model.

To further validate our model, we decided (i) to rationally tune the secretion efficiency of screened SPs, and (ii) to *in silico* screen a library of pseudo-randomly designed SPs for elements directing high secretion levels. Accordingly, from the previously screened SPs, we selected 30 sequences that poorly (Group 1) and 30 that highly directed secretion (Group 2), and we manually modified their nucleotide and amino acidic sequences to modulate their efficiency (Supplementary Table 5). An interactive exploration of original and edited SPs is possible through the File S2 (original) and File S3 (edited). Additionally, we generated 4,903 pseudo-randomly designed SPs, predicted their secretion efficiency, and picked 32 amongst the potentially best-performing SPs (average WA of selected SPs is 2.64) to be tested (Group 3). Out of these 92 SPs selected for model validation, 39 of the manually modified (Group 1 and 2) and 21 of the newly designed SPs (Group 3) were successfully cloned and tested for amylase activity in the MTP assay, showing substantial difference compared to their original counterparts (Groups 1 and 2) and very effective secretion (Group 3), respectively (Figure 3b). Remarkably, out of the 21 *in silico* pseudo-randomly designed SPs (Group 3), 5 showed a secretion efficiency higher than AmyQ with its native SP.

In a final effort to further validate the quality of the model in predicting SP efficiency in directing AmyQ secretion (i.e. WA value), we selected 5 SPs from each of the three groups and analysed their behaviour in the



	VC_2	VC_6	VC_8	VC_22	VC_23	VC_32	VC_37	VC_40	VC_52	VC_57	VC_60	VC_61	VC_62	VC_63	VC_68
○	7.88	7.44	7.71	8.18	5.85	1.01	1.48	1.02	1.17	1.64	N.A.	N.A.	N.A.	N.A.	N.A.
▽	7.59	7.42	6.54	7.28	7.08	1.15	1.48	1.14	1.23	1.79	N.A.	N.A.	N.A.	N.A.	N.A.
△	7.59	7.06	8.70	8.87	8.13	2.05	1.92	1.91	1.32	2.93	N.A.	N.A.	N.A.	N.A.	N.A.
▼	4.36	3.18	3.47	2.84	3.82	6.27	6.33	5.84	4.46	5.53	2.79	2.80	2.49	2.75	2.62
▲	5.28	2.67	7.70	2.17	4.10	5.96	8.05	6.07	4.33	6.98	2.13	2.70	2.96	4.49	1.66

Figure 3. Assay and model validation. (a) Comparison of the NLR- and MTP-based amylase assays for a random selection of 95 clones from the SP-library. The abscissa marks results from the MTP assay, with a value of 1 for the efficiency of the native SP of AmyQ; the ordinate marks weighted average (WA) values from the NLR assay. Of note, data points with a WA between 5 and 10 could not be measured with the standard MTP assay due to its low sensitivity; an optimized MTP assay and a hydrolysis test on starch agar plates was performed for the poorly secreting variants (Supplementary Figure 3 and Supplementary Table 3). Error bars represent the standard error of the mean over two replicates for the MTP-based assay; error bars are not represented for the NLR-based assay because their size was comparable to that of the marker. The dashed red regression trend line is based on all 95 data points (i.e. including those that could not be measured in the MTP-assay). **(b)** Bar plot showing amylase activity measured with the MTP assay for 60 SPs selected for model validation before (white bars, "ORIG") and after (black bars, "EDIT") editing. The efficiency of the native SP of AmyQ is equal to 1. Background shading distinguishes the 3 groups of SPs selected from the SP library: light yellow for originally poorly secreting SPs (Group 1), yellow for originally highly efficient SPs (Group 2), and dark yellow for pseudo-randomly designed SPs (Group 3). (* = not applicable, as the pseudo-randomly designed SPs were not present in the original SP-library). **(c)** and **(d)** Dot plots showing result of NLR-based analyses for *B. subtilis* strains secreting AmyQ with SPs VC_22 (from Group 1, whose secretion efficiency was improved) and VC_40 (from Group 2, whose secretion efficiency was reduced), respectively. White triangles indicate NLRs harbouring strains secreting AmyQ with the original variant from the SP-library, while black triangles indicate NLRs with strains secreting AmyQ with the modified SPs. In the background, the 10 different bins are indicated using the same colour code as in Figure 1b. Each dot plot was generated incubating in the same vessel (i.e. co-cultured) the library, the 'ORIG' and 'EDIT' version of the variant (10:1:1 ratio); the three populations could be differentiated based on the addition of diverse amounts of a blue fluorescently-labelled polymer. The overall lower green fluorescence of the library in Figure 3d, and the consequent reduction of the width of the bins, is due to the abundancy of NLRs harbouring a highly active variant (i.e. original VC_40) in the same vessel. As AmyQ is not retained in the NLRs, once the fluorescein-starch in the hydrogel bead is degraded, the protein diffuses among NLRs, reducing the overall fluorescence of the beads. Nonetheless, relative differences among the variants were detected in the defined time window. **(e)** Summary plot of the model validation. For each of the 15 selected SPs, 5 data points are shown: open symbols indicate the original variant from the SP-library, while black symbols designate the engineered SP derivative. The open circles mark the WA measured in the original NLR-based screening, which was used to build the model; triangles pointing downward denote WA values predicted by the model, whereas triangles pointing upward denote WA values actually measured during model validation. The data show that the differences between predicted and measured WA values mostly fall within the 10% range, highlighting the reliability of the activity assay and the model. Groups of SPs are highlighted in different shades of yellow as in panel B. Below the graph, all the plotted values are listed to allow a more detailed comparison.

NLR-based amylase activity assay. Dot plots of two variants (i.e. VC_22 and VC_40) with the respective original ('ORIG') and manually edited ('EDIT') versions are shown in Figures 3c and 3d, respectively and highlight the clear shift in secretion efficiency (Figure 3b) for the two versions (from low to high for VC_22, and high to low for VC_40). Figure 3e summarizes the WAs of the 15 selected SPs and compares them with the WAs obtained at each step of the workflow (i.e. library screening, model predictions and validation). Remarkably, 11 of the tested SPs fell within one unit of difference (i.e. +/- 1 WA) from the predicted value, implying that the proposed workflow is indeed a powerful tool to quantify and engineer secretion efficiency of SPs. To our knowledge, this is an unprecedented prediction accuracy, where part of the variation in the estimated values could relate to small differences in the culturing conditions between library screening and validation. Also, an increased frequency of high-performing variants during validation may have slightly shortened the optimal time-window of the kinetic assay. Additionally, part of the discrepancies between predicted and measured values can be explained by a few limitations of the library design. Despite having an extensive experimental space for each single feature, the library was not designed to be fully combinatorial and did, thus, not exhaust the full design space. This potential limitation of the current approach may be overcome in future studies by using a fractional factorial design[32,33] to ameliorate the design space, e.g. combining regions with differently modified features rather than editing one at a time. A particularly exciting outlook for future studies is to expand our present approach to different SP-protein combinations, including features describing the mature protein along with possible interactions; because secretion-relevant features are embedded not only in the SP, but also in the mature protein sequence[34,35].

Altogether, we conclude that our presented approach can detect and explain the relevant SP features influencing the efficiency of protein secretion. It thus sets the stage for *in silico* tuning and *de novo* design of SPs. Although we limited our present study to one protein, the workflow can easily be extended to other industrially or biomedically relevant POIs by applying different enzymatic assays[36] and novel HT analytical systems[37,38]. In fact, *in silico* SP design based on our trained model already proved very effective in the present proof-of-principle study since the best predicted SPs turned out to direct high-level secretion.

For the future, we advocate an iteration of the here proposed DBTL cycle to obtain further insights into the general features that influence protein secretion, especially through the screening of different POIs, fused to the same set of SPs. Datasets thus obtained will further improve the generalizability and reliability on prediction and design of SPs directing high secretion levels. As a result, far smaller numbers of SPs will be screened, or SPs sequences will directly be designed *in silico*, for instance with the presented pseudo-random approach, or by exploiting a novel ML-based tool for SP generation[39]. We are therefore confident that, with less experimental testing, our approach will deliver more accurate, better tunable, and highly productive protein secretion systems.

Acknowledgments

This work was supported by the European Union's Horizon 2020 Program, Marie Skłodowska-Curie Actions (MSCA), under REA grant agreement no. 642836. We would like to thank Nivitha Punniyamoorthy and Steven Schmitt for the technical support provided during the long hours spent at the Biosorter. We would also like to thank Irsan Kooi, Marcel Hillebrand, Rianne van der Hoek, and Ana Bulović for the fruitful discussions.

Conflict of interests

The authors declare no competing interests. However, M.Hen., P.Z., and T.v.R. are Scientists at DSM B.V.; while V.D., R.P., and A.M are Scientists at FGen GmbH.

Data availability

The data that support the findings of this study is available as Supplementary Information.

Code availability

The ML model and SHAP data are available at https://github.com/grassoste/SP-secretion-efficiency-Jupyter-notebook. Additionally, also an interactive Jupyter notebook to fully explore the model is available in the same repository.

Author contributions

J.M.vD., M.Hel., S.P., A.M., and T.v.R. conceived the project. S.G. and P.Z. designed the libraries. S.G. designed and performed cloning and transformation. V.D. and R.P. designed and optimized the NLR-based enzymatic assay. S.G. and V.D. performed the screening assay. S.G. prepared the sequencing library and analyzed the NGS data. S.G. and M.Hen. generated and explained the ML model. V.D. validated the assay and the model. S.G. and V.D. drafted the manuscript. S.P., J.Mv.D., A.M., and T.v.R. supervised the project and revised the manuscript.

References

1. Goosens, V. J., Monteferrante, C. G. & Van Dijl, J. M. The Tat system of Gram-positive bacteria. *Biochim. Biophys. Acta - Mol. Cell Res.* **1843**, 1698–1706 (2014).
2. Desvaux, M., Hébraud, M., Talon, R. & Henderson, I. R. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol.* **17**, 139–145 (2009).
3. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.* **97**, 422–441 (2018).
4. Denks, K. *et al.* The Sec translocon mediated protein transport in prokaryotes and eukaryotes. *Mol. Membr. Biol.* **31**, 58–84 (2014).
5. Anné, J., Economou, A. & Bernaerts, K. Protein Secretion in Gram-Positive Bacteria: From Multiple Pathways to Biotechnology. in *Current Topics in Microbiology and Immunology* vol. 404 267–308 (Springer Verlag, 2016).
6. Demain, A. L. & Vaishnav, P. Production of recombinant proteins by microbes and higher organisms. *Biotechnol. Adv.* **27**, 297–306 (2009).
7. Ferrer-Miralles, N. & Villaverde, A. Bacterial cell factories for recombinant protein production; expanding the catalogue. *Microb. Cell Fact.* **12**, 113 (2013).
8. van Dijl, J. M. & Hecker, M. Bacillus subtilis: from soil bacterium to super-secreting cell factory. *Microb. Cell Fact.* **12**, 3 (2013).
9. Harwood, C. R. & Cranenburgh, R. Bacillus protein secretion: an unfolding story. *Trends Microbiol.* **16**, 73–9 (2008).
10. Tjalsma, H., Bolhuis, A., Jongbloed, J. D., Bron, S. & van Dijl, J. M. Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. *Microbiol. Mol. Biol. Rev.* **64**, 515–47 (2000).
11. Tjalsma, H. & Van Dijl, J. M. Proteomics-based consensus prediction of protein retention in a bacterial membrane. *Proteomics* **5**, 4472–4482 (2005).
12. Peng, C. *et al.* Factors influencing recombinant protein secretion efficiency in gram-positive bacteria: Signal peptide and beyond. *Front. Bioeng. Biotechnol.* **7**, (2019).
13. Nielsen, H., Tsirigos, K. D., Brunak, S. & von Heijne, G. A Brief History of Protein Sorting Prediction. *Protein J.* **38**, 200–216 (2019).
14. Brockmeier, U. *et al.* Systematic Screening of All Signal Peptides from Bacillus subtilis: A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-positive Bacteria. *J. Mol. Biol.* **362**, 393–402 (2006).
15. Degering, C. *et al.* Optimization of protease secretion in bacillus subtilis and bacillus licheniformis by screening of homologousand heterologous signal peptides. *Appl. Environ. Microbiol.* **76**, 6370–6376 (2010).
16. Mathiesen, G. *et al.* Genome-wide analysis of signal peptide functionality in Lactobacillus plantarum WCFS1. *BMC Genomics* **10**, 425 (2009).
17. Yu, M. K. *et al.* Visible Machine Learning for Biomedicine. *Cell* **173**, 1562–1565 (2018).
18. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* (2017).
19. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
20. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
21. Lundberg, S. M., Erion, G. G. & Lee, S.-I. Consistent Individualized Feature Attribution for Tree

Ensembles. *ArXiv* **cs.LG**, (2018). Preprint available at: https://arxiv.org/abs/1802.03888

22. Hall, P., Gill, N. & Schmidt, N. Proposed Guidelines for the Responsible Use of Explainable Machine Learning. *ArXiv* **stat.ML**, (2019). Preprint available at: https://arxiv.org/abs/1906.03533
23. Meyer, A. *et al.* Optimization of a whole-cell biocatalyst by employing genetically encoded product sensors inside nanolitre reactors. *Nat. Chem.* **7**, 673–678 (2015).
24. Walser, M., Leibundgut, R. M., Pellaux, R., Panke, S. & Held, M. Isolation of monoclonal microcarriers colonized by fluorescent *E. coli*. *Cytom. Part A* **73A**, 788–798 (2008).
25. Briggs, D. E. Gel-diffiusion method for the assay of α-amylase. *J. Inst. Brew.* **68**, 27–32 (1962).
26. Strahl, H., Bürmann, F. & Hamoen, L. W. The actin homologue MreB organizes the bacterial cell membrane. *Nat. Commun.* **5**, 1–11 (2014).
27. Zeng, W., Guo, L., Xu, S., Chen, J. & Zhou, J. High-Throughput Screening Technology in Industrial Biotechnology. *Trends Biotechnol.* **38**, 888–906 (2020).
28. Tjalsma, H. *et al.* Proteomics of Protein Secretion by Bacillus subtilis : Separating the 'Secrets' of the Secretome. *Microbiol. Mol. Biol. Rev. 2* **68**, 207–233 (2004).
29. Choo, K. H. & Ranganathan, S. Flanking signal and mature peptide residues influence signal peptide cleavage. *BMC Bioinformatics* **9 Suppl 12**, S15 (2008).
30. Zanen, G. *et al.* Signal peptide hydrophobicity is critical for early stages in protein export by Bacillus subtilis. *FEBS J.* **272**, 4617–4630 (2005).
31. Dyrløv Bendtsen, J., Nielsen, H., von Heijne, G. & Brunak, S. Improved Prediction of Signal Peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
32. Naseri, G. & Koffas, M. A. G. Application of combinatorial optimization strategies in synthetic biology. *Nat. Commun.* **11**, 2446 (2020).
33. Gunst, R. F. & Mason, R. L. Fractional factorial design. *Wiley Interdiscip. Rev. Comput. Stat.* **1**, 234–244 (2009).
34. Chatzi, K. E. *et al.* Preprotein mature domains contain translocase targeting signals that are essential for secretion. **216**, (2017).
35. Smets, D., Loos, M. S., Karamanou, S. & Economou, A. Protein Transport Across the Bacterial Plasma Membrane by the Sec Pathway. *Protein J.* **38**, 262–273 (2019).
36. Reymond, J.-L., Fluxà, V. S. & Maillard, N. Enzyme assays. *Chem. Commun.* 34–46 (2008) doi:10.1039/B813732C.
37. Markel, U. *et al.* Advances in ultrahigh-throughput screening for directed enzyme evolution. *Chem. Soc. Rev.* **49**, 233–262 (2020).
38. Bunzel, H. A., Garrabou, X., Pott, M. & Hilvert, D. Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. *Curr. Opin. Struct. Biol.* **48**, 149–156 (2018).
39. Wu, Z. *et al.* Signal Peptides Generated by Attention-Based Neural Networks. *ACS Synth. Biol.* **9**, 2154–2161 (2020).

Online Methods

Library design

To identify the most relevant physicochemical features influencing the secretion efficiency directed by signal peptides (SPs), we designed a SP mutant library starting from 134 sequences (Supplementary Table 1) of known or highly probable *Bacillus subtilis* wild-type SPs. These SPs were initially selected based on literature[1] and via predictions by various computational tools (SignalP4.1[2], SignalP3.0[3], Phobius[4]). Next, the selected SPs were manually curated to remove false positives, knowing the final localization of their cognate native protein. As a point of novelty, we considered the SP sequences both as a single sequence (i.e. the whole SP) and as the juxtaposition of 4 separate parts, namely: the canonical N-, H-, and C-regions, plus a region referred to as 'after cleavage' (Ac-region), which consists of the first 3 amino acid residues (AAs) after the expected signal peptidase cleavage site. The Phobius tool for SP predictions was used to determine the boundaries of the 4 regions constituting each SP, still with partial manual curation based on evidence from literature. After defining the 4 regions for each SP, physicochemical properties were calculated for each region independently as well as for the complete SP. The 227 calculated properties are listed in Supplementary Table 2 while the respective methods of calculation and further explanations are reported in Supplementary Table 1.

From each of these 134 SPs, 94 mutant sub-libraries of 134 elements each were created. In each sub-library only one feature at a time was edited, while modifications to other interdependent features (e.g. the charge of an AA sequence affects also its isoelectric point and hydrophobicity) were minimized. Edited features at the AA level were hydrophobicity, charge, and length; edited features at the nucleotide level were codon usage and RNA secondary structures. The full list of varied features is presented in Supplementary Table 2. For each selected feature, multiple target levels (usually 4 or 5) were chosen. The rationale for selecting target levels was to allow for some expansion of the investigated design space without diverging too much from the biologically meaningful space of the wild-type SPs. The resulting SP-library was composed of the 94 sub-libraries and included a total of 11,643 unique sequences, which are presented in Supplementary Table 1.

pSG01 plasmid construction

The plasmid pSG01 (see Supplementary Figure 10, and Supplementary Table 6 for the full plasmids list) was developed within this study in order to be used as a chromosomal integration vector for expression of the SP-library. To this end, the previously constructed genome-integrating vector pCS75[5] (Supplementary Table 6) was cleaved with *Pme*I and *Eag*I (NEB), the resulting fragments were separated on a 0.8% agarose gel, and the 7.8 kbp band, delimited by two regions homologous to the *B. subtilis amyE* gene, was excised and purified with the QIAquick Gel Extraction kit (Qiagen). The DNA sequence encoding the AmyQ mature protein (P00692) (i.e. without its SP) was ordered as a single gBlock G1 (Integrated DNA Technologies, Inc.) (see Supplementary Table 6 for full nucleotide sequence), amplified with primers P1 and P2 (see Supplementary Table 6 for a full primer list), digested with the same restriction enzymes as the vector, and purified with the DNA Clean & Concentrator-25 kit (ZymoSearch). The two DNA fragments were ligated, and the ligation mix was directly used to transform 10-beta competent *Escherichia coli* cells (NEB), to amplify pSG01. The resulting plasmid was verified and used to transform *dam*$^-$/*dcm*$^-$ competent *E. coli* cells (NEB), from which demethylated pSG01 was obtained for all downstream applications to increase the efficiency of *B. subtilis* transformation[6]. Notably, 5' to the SP-less *amyQ* gene, plasmid pSG01 contains two *BsmBI* (a type IIS restriction enzyme) restriction sites at 11 nt distance, which are oppositely oriented so that cleavage occurs upstream of each restriction enzyme recognition sequences, thus allowing for scar-less insertion of properly oriented DNA fragments. Moreover, this feature allows for the insertion of multiple DNA fragments in one step. After transformation of *B. subtilis*, pSG01 will integrate into the *amyE* gene, thereby disrupting the main source of amylase activity in *B. subtilis*.

Expression strains and cloning of the library

B. subtilis strain DB104[7], which lacks two major extracellular proteases, was selected to produce the library of designed SPs fused to AmyQ.

To obtain the final SP-library, pSG01 was endowed with two DNA fragments, using the two *BsmBI* restriction sites upstream of the SP-less *amyQ* gene: one fragment contained the P_{veg} promoter[8], the native mRNA stabilizer of *cotG*[9], and a strong RBS from the *pre(mob)* gene of pUB110[10], obtained as a single gBlock G2 (Integrated DNA Technologies, Inc.; Supplementary Table 6); the other fragment coded for one of the 11,643 designed SPs (ordered as an oligo pool from Twist Bioscience). Both fragments were designed to be amplified with P1 and P2 primers (Supplementary Table 6) and to present two terminal *BsmB*I cleavage sites generating complementary sticky ends to the vector for sequential assembly. Cloning was carried out using the StarGate[11] methodology and the resulting construct constitutively expressed the gene coding for the mature AmyQ fused at the N-terminus with one of the 11,643 designed SPs. A total of 3 mL StarGate reaction was mixed with 63 mL of competent *B. subtilis* DB104 that has been prepared using a modified Spizizen protocol[5]. After 1 h of recovery at 37°C and 250 rpm, cells were plated on 62 Q-trays (Nunc™ Square BioAssay Dishes product n. 240835, ThermoFisher) each containing 200 mL of 2xPY medium (16 g/L peptone, 10 g/L yeast extract, 5 g/L NaCl) supplemented with 15 g/L agar and 300 µg/mL spectinomycin. After cell plating, the Q-trays were incubated at 30°C for 20 h.

The total number of grown colonies was estimated using a QPix 450 (Molecular Devices) automated microbial screening system. Two rounds of transformation were performed in order to obtain approximately 160,000 colonies, corresponding to a 10X coverage of the SP-library, and estimating 10% of clones containing pSG01 without inserts (data not shown). Plates were scraped to collect all colonies and rinsed with 2xPY. The collected cells were then transferred to several 50 mL Falcon tubes, mixed, and concentrated by centrifugation at 3,000xg for 5 min. The pellets were resuspended in 2xPY, the cell suspensions were pooled, thoroughly mixed, and supplemented with glycerol to a final concentration of 10% (v/v). The glycerol stock was aliquoted, snap frozen, and stored at -80°C. The cell concentration in the glycerol stocks, as determined by the optical density at 600 nm, was approximately $5.8*10^9$ cells/mL.

Substrate preparation for NLR-based amylase assay

Dry corn starch (Sigma Aldrich, S9679) was re-suspended in 90/10 DMSO/water (v/v) to a final concentration of 2% (w/v), boiled for 30 min and allowed to cool to room temperature. An aliquot of 100 mL of the prepared solution was basified with 1 M NaOH until it reached a pH ≥ 9, then mixed with 1 mL of the reactive dye 5-([4,6-dichlorotriazin-2-yl]amino)fluorescein hydrochloride (DTAF) (Sigma Aldrich), previously dissolved in DMSO (20 mg/mL). After 1 h incubation at room temperature, the solution was neutralized with glacial acetic acid to stop the reaction, and the fluorescein-starch was precipitated with ethanol to remove the remaining free dye. The precipitated starch was resuspended in DMSO and subsequently ground with glass beads at 30 Hz for 20 min (Retsch). The resulting fluorescein-starch preparation was stored at 4°C and used as the substrate employed to monitor amylase activity within the nanoliter reactor (NLR)-based assay described below.

Cultivation of strains in NLRs

NLRs were synthesized starting from a mix of bacterial glycerol stocks, fluorescein-starch and sodium alginate, which was processed through a laminar jet break-up encapsulator (Nisco Engineering) to generate a monodisperse bead population. To prepare the mix, 200 µL of fluorescein-starch (4% w/v in DMSO) were diluted in 2 mL of resuspension medium (4 g/L yeast extract, 1 g/L tryptone, 20 mM TRIS pH 7) and added to 16 mL of sodium alginate 2.5% (w/v) aqueous solution. The number of bacterial cells to be included was

defined to achieve an average occupation of 0.3 cells per NLR. To this end, the corresponding volume of the bacterial glycerol stock was added to the resuspension medium to reach a final volume of 2 mL, which was then mixed with the fluorescein-starch alginate preparation.

For NLR formation, the encapsulator was equipped with a 150 µm nozzle, and operated with a flow rate of 3.3 mL/min and a frequency of 650 Hz[12]. This delivered NLRs with an average diameter of 500 µm (corresponding to a volume of approximately 65 nL). NLRs were allowed to harden for 15 min in 100 mM aqueous $CaCl_2$, then isolated using a cell strainer (100 µm mesh size, Falcon, Becton Dickinson) and washed once with 10 mM aqueous $CaCl_2$. NLRs were transferred into growth medium (4 g/L yeast extract, 1 g/L tryptone, 20 mM TRIS pH 7, 4 mM $CaCl_2$, 300 µg/mL spectinomycin) with 0.5% (v/v) amylopectin to a final concentration of 100 g wet NLRs/L in Erlenmeyer flasks. The reactors were incubated in a shaker (150 rpm, room temperature) for approximately 13 h to allow cells to grow into microcolonies. NLRs were then recovered and washed twice with screening buffer (10 mM $CaCl_2$, 10 mM TRIS pH 8). During each wash, the beads were allowed to sediment in a 50 mL Falcon tube, the supernatant discarded, and buffer added to achieve a concentration of 12.5 g of wet NLRs/L. Prior to screening, 40 µL of Nile red (Chemodex) (1 g/L in 90/10 DMSO/water, v/v) were added for every gram of wet NLRs to fluorescently stain the cells. The NLRs were incubated for 20 min under gentle shaking, washed once more with the screening buffer to remove surplus dye, and then subjected to flow cytometry and microscopic analysis. Bright-field and fluorescence microscopy images were recorded using an Axio Observer II with an AxioCam MR3 camera (Carl Zeiss Microscopy) to control for proper NLR synthesis and cell growth. For a detailed description of the flow cytometry analysis, see the section below.

If alginate beads with known SPs variants needed to be incubated together in the same vessel (to guarantee identical incubation conditions) and differentiated later in the flow cytometry analysis, the NLRs were synthesized with different concentrations of Pacific-blue (Ex 410 nm, Em 455 nm) labelled amino dextran (AD). Two concentrations, corresponding to 12 and 2.4 µL of the Pacific-blue AD stock solution (20 mg/mL in 0.2 M sodium bicarbonate, pH 8.3) per mL of fluorescein-starch alginate, were added. This polymer is not a substrate for AmyQ (data not shown) and does not interfere with the recording of fluorescein-based fluorescence (Ex 492 nm, Em 516 nm). Instead, the Pacific-blue content can be read out in the violet spectrum. Conjugation of the dye to the polymer was achieved by adding 5 mg of the amine-reactive Pacific Blue succinimidyl ester (ThermoFisher) to a solution of 20 mg AD (Fina Biosolutions) per mL of 0.2 M sodium bicarbonate (pH 8.3). The reaction was incubated for 6 h at room temperature. Then TRIS pH 7 was added to a final concentration of 50 mM to stop the reaction, and the solution was aliquoted and frozen.

Throughout the study, different *B. subtilis* strains, all generated in the same fashion and with the same vector, were analyzed using the NLR-based amylase assay. These included: 1) *B. subtilis* producing AmyQ with its native signal peptide (positive control, PC), 2) *B. subtilis* carrying the empty vector, without an inserted signal peptide (negative control, NC), 3) *B. subtilis* transformed with the SP-library, fused to AmyQ, or 4) *B. subtilis* producing AmyQ with SP variants with defined modification. The PC (1) and the NC (2) strains, and two variants producing AmyQ with known SPs were used to estimate the dynamic range and sensitivity of the NLR-based amylase activity assay (Supplementary Figure 1 and Supplementary Figure 2). 15 strains producing AmyQ with SP variants with defined modification (4) were encapsulated and used to validate both the NLR-based screening assay and the model.

NLR-based screening

The NLR-based screening of the clones carrying the SP-library, was performed with a large particle flow cytometer, which allowed to read out the amount of starch, of cells, and, if applicable, of amino dextran in each NLR, based on different fluorescence signals. Specifically, we used a BioSorter (Union Biometrica) to record for each NLR green (excitation laser 488 nm, beam splitter DM 562, emission filter BP 510/23 nm), red (excitation

laser 561 nm, TR mirror, emission filter BP 615/24 nm) and violet fluorescence (excitation laser 405 nm, beam splitter DM 495, emission filter BP 445/40 nm). Each screening round was performed in two sequential steps. During the first step, all events were analyzed in bulk mode, at a maximum of 90 Hz, and NLRs with a positive red fluorescence (peak height, i.e. presence of colonies stained with Nile red) were sorted into a 50 mL Falcon tube, containing 5 mL of screening buffer. The isolated population represented approximately 20% of all the NLRs, in agreement with the occupation estimated from the cell concentration in the glycerol stocks. Prior to the second step, the values of green and red fluorescence of each sorted NLR were graphically visualized using the FlowPilot software provided by the BioSorter manufacturer. The graph was then used to divide all events in 10 bins based on the ratio between green fluorescence (area, representative of amylase activity and secretion levels) and red fluorescence (peak height, representative of total biomass). The bin width was thus adjusted to have 10% of the events sorted in step 1, in each bin. For the second step, sorted NLRs were run through the Biosorter ten consecutive times, every time isolating in bulk mode the events falling in one bin. In particular, the sorting started from the bin with the lowest green to red ratio (i.e. highest secretion/biomass ratio), bin 1, and then moved progressively to bins with higher green/red ratios. The screening analysis was repeated 9 times until the number of occupied NLRs (i.e. positive red fluorescence) reached 160,000, to ensure a 10x coverage of the SP-library. Additionally, after cell encapsulation and growth in the NLRs, 53,588 occupied NLRs were sorted in 3 rounds, without performing any further binning, and treated separately. This sample, named hereafter 'occupation control', was processed and sequenced with the 10 bins, and later used to gather information about the library coverage and the *B. subtilis* population at this step of the workflow.

To recover the NLR-embedded cells, binned samples were incubated for 10 min under gentle shaking with 2xTY medium (16 g/L tryptone, 10 g/L yeast extract, 5 g/L NaCl) supplemented with potassium phosphate buffer (pH 7) to a final concentration of 0.2 M, at which point full dissolution of the cross-linked calcium alginate had been achieved. Bacterial cells were pelleted by centrifugation (4,000xg, 30 min), the supernatant was discarded, and the pellet stored at -80°C.

Genomic DNA extraction and NGS library preparation

Samples from the 10 bins, the occupation control, and the initial glycerol stock were thawed on ice, centrifuged for 1 min at 16,000xg, and the supernatant was discarded. Afterwards cells were resuspended in 0.85% (w/v) aqueous NaCl, supplemented with 250 µg/mL of RNase A (Macherey-Nagel) and 0.5 mg/mL of lysing enzymes from *Trichoderma harzianum* (Sigma-Aldrich, L4142). After incubating for 10 min at 37°C, EDTA and SDS were added to final concentrations of 15 mM and 1.2%, respectively. Samples were vortexed thoroughly, ammonium acetate was added to a concentration of 2.5 M, and then samples were vortexed again. Precipitated proteins were pelleted by centrifugation at 22,000xg for 15 min at 4°C. The supernatant was transferred to a fresh reaction cup, supplemented with an equal volume of 2-propanol and gently mixed. DNA was then pelleted by centrifugation at 22,000xg for 40 min at 4°C. The supernatant was discarded and the pellet was washed twice with ice-cold 70% ethanol, dried, and resuspended in 10 mM Tris-HCl pH 7.5.

Each sample was then amplified by PCR, using Phusion polymerase (NEB), with primers P3-P15 (Supplementary Table 6) that anneal immediately up- and downstream of the inserted SP sequence in pSG01, adding barcodes to identify the sample (primers P3-P15 in Supplementary Table 6, containing Illumina Nextera tagmentation adapters and, in each forward primer, a specific barcode). PCR products were then purified and recovered in milliQ water. Amplicons were analyzed with a bioanalyzer (LabChip GXII, Caliper Life Sciences) using a 5K HT DNA chip, to check size and concentration of the fragment. The 12 PCRs products, corresponding to the 10 bins and the two controls, were pooled and sequenced as a Nextera library (Illumina) by the company BaseClear B.V. (Leiden, NL) on a NovaSeq machine (Illumina) in paired ends, for a total of 26,175,197 2x150bp reads. For both forward and reverse raw reads, the Phred scores had an average of 36 and a median of 37.

Reads pre-processing and mapping

The software FastQC version 0.11.8[13] was used for quality inspection of the sequencing data. First, possible adapters were removed from the 3' end of the reads (read-trough adapters), since they could confound the merging process when the read length and insert size are comparable. To this end, the software NGmerge[14] version 0.2dev was used in "adapter removal" mode, with 0 mismatches allowed. Sequences were thus merged into longer pseudoreads using PEAR[15] version 0.9.11, with a minimum overlap of 5nt and a p-value of 0.001. This yielded 26,105,901 pairs of reads (99.735% of the total reads) to be merged, with the remaining reads unassembled and no read discarded. Pseudoreads were then sorted in the 10 bins and the two controls, based on the respective barcodes, using the 'fastx_barcode_splitter.pl' script from FASTX-Toolkit[16] looking only at the 5' ('*--bol*' option) and allowing only 1 mismatch. This resulted in 25,980,025 (99.254% of the total reads) demultiplexed pseudoreads. Any remaining adapter (including the barcode) at both 5' and 3' of the assembled reads was removed using cutadapt[17] version 2.3, without any read loss. The obtained pseudoreads were then mapped to the reference sequences (i.e. the designed SPs) using BBMap[18] version 37.93, with '*perfectmode*' activated; and behavior for ambiguously mapped reads was set to 'best alignment' (Supplementary Table 4). Occurrences for each bin and both controls were counted for each of the designed sequences and the resulting frequency table was later used for model construction (Supplementary Table 1).

Data preprocessing, feature extraction, model construction and interpretation

To identify the possible influence of investigated features on protein secretion, we decided to train a simple machine learning model. This procedure, combined with an interpretation of the model, would allow us to obtain a predictive model that could yield important mechanistic insights into the features determining the secretion efficiency of different SPs.

First of all, sequences with low abundance, corresponding to less than 255 reads in the most populated bin, were discarded. This resulted in 4,421 informative SPs, which were used to train and test the model. As a different number of NLRs was collected for each bin, the occurrence of reads was normalized across bins so that they contained the same number of NLRs. To score SPs, we assumed that bins were equidistant and each bin had an average value corresponding to its number. A weighted average (WA), i.e. the summation of bin values weighted on the relative frequencies of reads, was calculated for each SP and used as a secretion score. The WA values of selected SPs could thus range from 1 (i.e. the best secreting SPs with all occurrences detected in bin 1) to 10 (i.e. the worst secreting SPs with all occurrences detected in bin 10).

From the 227 calculated features, 22 were discarded because they showed no variation either in the designed SP-library or in the informative SPs dataset, which was a subset of the designed library (Supplementary Table 2). Furthermore, it was decided to minimize the number of features with a correlation coefficient higher than 0.7 to avoid a spread of importance, as attributed by the model, among them. Thus, out of the initial 227 features, 96 were retained, while 110 features, with correlation coefficients higher than 0.7, were selected for clustering. Clustering was carried out through affinity propagation[19] using the scikit-learn[20] python package with standard parameters. Notably, affinity propagation was selected as the clustering algorithm, because of its intrinsic capability of inferring the total number of clusters. This resulted in 14 clusters of correlating features, out of which 22 features were selected and added back to the feature set. Specifically, for 12 of the clusters the centroid was selected, for 1 cluster the centroid and an additional feature were selected, and for the last cluster with lowly correlating features all of the 7 features were included. Altogether, this procedure resulted in a total of 116 features, to which 40 Boolean dummy variables were added that describe the AAs in positions -3 and -1, respective to the signal peptidase cleavage site. This resulted in the final selection of 156 features describing the selected SPs (Supplementary Table 2). In order to verify that the selected features were relevant (i.e. provided variance) within the datasets, and thus needed to train the model, a principal

component analysis was performed on: (i) the 134 wild-type known SPs, (ii) the set of 11,643 unique SP (i.e. the SP-library), and (iii) the set of 4,421 informative SPs (Supplementary Figure 4).

To construct the model, the matrix of 4,421 informative SPs and 156 features was used as the independent variable, while the array of 4,421 WA values was used as the dependent variable. These matrices were split with the Kennard-Stone algorithm[21,22] into a training set and a test set of 3,095 and 1,326 informative SPs, respectively. From available models, a Random Forest[23,24] (RF) Regressor model from scikit-learn[20] was implemented. To identify the best hyperparameters a 5-time cross-validation grid search was performed using the training set. From the 15,435 tested combinations of hyperparameters, the following set of hyperparameters was selected, which balance the predictive power and the size of the model: max depth 25, max features 156, min samples leaf 0.0001 of the training set, min samples split 0.001 of the training set, and estimators 75. The model was subsequently evaluated on the test set, and scored calculating the mean squared error between measured and predicted values (Supplementary Figure 5).

The RF model was analyzed to gain mechanistic insights and an explanation of the model itself. For this task, the TreeSHAP method from the SHAP (SHapley Additive exPlanation)[25–27] package was used since, being based on Shapley values, it is advantageous in terms of consistency, allows for a more reliable comparison of feature attribution values, and allows users to understand the model explanation. A further advantage of SHAP is that it includes both local and global explanations, thereby providing explainability for both the whole dataset and the individual SPs. Nevertheless, it should be emphasized that SHAP only provides an explanation of the model based on the contribution of individual features to the final output. SHAP does not necessarily uncover the causal relationships between individual SP features and the actual protein secretion efficiency as displayed by a bacterial cell. Furthermore, is noteworthy that SHAP provides the possibility to determine the type of relationship between each individual feature and the predicted output, and to determine second order interactions that occur between features.

Assay validation

To show both the reliability of the NLR-based amylase activity assay and to assess the correctness of the model, we used two orthogonal procedures to measure the amylase activity from strains producing AmyQ with selected SPs: a commercial assay in 96-wells microtiter plates (MTPs) and a hydrolysis test on starch agar plates[30]. For the MTP assay bacteria were precultured overnight into 300 µL 2xPY supplemented with 300 µg/mL spectinomycin, subsequently diluted 100-fold in the same medium and grown for 7.5 h at 37°C and 250 rpm. Aliquots of the different cultures were collected, the cells were pelleted, and 9 µL of the supernatants were mixed with 9 µL of Ceralpha reagent (Megazymes). The reaction mixtures were incubated for 20 min (standard version) or 90 min (sensitive version) at room temperature on a shaker (1,000 rpm) and then the reactions were stopped through the addition of 200 µL of 1% (w/v) 2-amino-2-(hydroxymethyl)propane-1,3-diol (Tris-base, pH 9). The amylase activity was then measured by monitoring the absorbance at 405 nm with a Tecan Infinite M200 Pro. Similar to the NLR-based assay, the optical density at 600 nm of the cultures was measured and used to normalize all samples for the biomass. Eventually, the OD-normalized absorption values were expressed relative to the amylase activity obtained from cultures that secreted AmyQ with its native SP (positive control, PC), defined as 100%, and to the amylase activity in the growth medium of a strain containing pSG01 without an inserted SP sequence (negative control, NC), defined as 0%.

For the hydrolysis test on starch agar plates, glycerol stocks of the selected variants were diluted 100-fold in 300 µL 2xPY supplemented with 300 µg/mL spectinomycin, and grown until they reached mid-exponential phase (6 h, 37°C, 250 rpm). An aliquot of 2 µL of the cell culture was spotted on 2xPY-agar plates supplemented with 300 µg/mL spectinomycin and 0.2% (w/v) potato starch (Sigma Aldrich). After overnight incubation at 37°C, the plates were flooded with Lugol's iodine (Carl Roth), which interacts with starch and generates a dark color. Where starch is degraded, a clear zone arises, e.g. around a colony, and the area of

this clear degradation zone is approximately proportional to the amount of amylase secreted[30].The standard MTP assay was used for initial validation of the NLR-based assay (Supplementary Figure 2) and the screening strategy. For that, 95 SP-AmyQ fusions, randomly picked from the 4,421 variants used to train and test the model were subjected to the MTP assay (see activity values in Supplementary Table 3 and Figure 3a). As 72 randomly picked variants showed no amylase activity, presumably due to too low secretion levels, the sensitive version of the MTP assay was applied, and AmyQ activities could be determined for 15 more clones.

The starch hydrolysis on plate was also applied for the partial validation of the NLR-based assay (Supplementary Figure 2) and if no amylase activity could be detected with the sensitive MTP assay (i.e. $Abs_{405} < 0.1$; Supplementary Figure 3).

Model validation

To further validate our model, small sets of SPs were manually edited to tune their predicted secretion levels. 30 SPs directing high-level secretion of AmyQ (i.e. 'good performers') were manually edited until the model predicted them to direct AmyQ secretion with poor efficiency (Group 1). Similarly, another 30 SPs directing low-level secretion of AmyQ were edited in order to improve their efficiency (Group 2) (see Supplementary Table 5 for full list of SPs).

As an additional validation approach, we generated pseudo-random SP AA sequences, with a home-made script, described as follows. Based on 134 sequences of known and highly probable SPs, 7 dictionaries were calculated that map each AA to its relative frequency: one for the N-region (excluding the initial Met), one for the H-region, one for the C-region except the last 3 residues, one for each -3, -2 and -1 position relative to the signal peptidase cleavage site, and the last one for the Ac-region (i.e. positions +1, +2 and +3 together). Using these values, 10,000 sequences for each region were generated as follows: for the N-region a Met was always placed in front of a stretch of 1 to 10 residues built on the frequency table; the H-region simply consisted of a stretch of 9 to 16 residues built on the frequency table; the C-region was built juxtaposing a stretch of 4 to 11 residues to the 3 single residues for positions -3,-2, and -1, each based on its frequency table; and for the N-terminus of the mature protein after signal peptidase cleavage, a stretch of 3 residues based on its frequencies was built. To minimize the occurrence of SPs with features too far from the distribution of the known, or probably representing wild-type *B. subtilis* SPs, the Kolmogorov-Smirnov[28,29] statistic test was applied, which compared the distribution of various features (data not shown). In case the number of similar distributions (considered as all those features with a calculated p-value above 0.1) was below a certain threshold, i.e. at least 21, 16, 18 and 17 features respectively for the N-, H-, C-, and Ac-regions, the batch of 10,000 sequences was discarded and the process was repeated. In such a way, the 4 designated regions (i.e. the N-, H-, C-, and Ac-regions) were built independently from each other and their possible combinations or interactions were not considered. When 10,000 sequences for each region were generated, they were juxtaposed to form a full SP. Then, sequences equal or longer than 33 AAs (calculating the length up to and including the Ac-region) were discarded, thus resulting in 4903 valid SPs. To generate the relative coding sequences, the AA sequences were retro-translated with an unambiguous dictionary where only the most frequent codon for each AA was present. Subsequently, having both a nucleotide and an AA sequence for each SP, the 156 features of the final model were calculated, and the respective SPs' secretion efficiency (i.e. the WA value) was predicted using the RF model. A set of 32 SPs (Group 3) was selected with the requirement to be predicted by our model as very good secretors (see Supplementary Table 5 for full list of SPs). Eventually, the different features for each SP were also analyzed and explained through SHAP.

For manually edited as well as pseudo-randomly designed SPs, the respective SP-encoding sequences were ordered (Twist Biosciences), cloned in pSG01 and used to transform *B. subtilis* DB104, following the same procedure as applied for the generation of the SP-library. For the 61 successfully constructed SP-AmyQ fusions, the amylase activity was monitored in the MTP assay as described above. Additionally, 15 of these

fusions (5 for each of the three groups) were further tested via the NLR-based amylase activity assay, to verify the model in the same experimental setup used to generate it (Figure 3).

References for Online Methods

1. Brockmeier, U. *et al.* Systematic Screening of All Signal Peptides from Bacillus subtilis: A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-positive Bacteria. *J. Mol. Biol.* **362**, 393–402 (2006).
2. Petersen, T. N., Brunak, S., Von Heijne, G. & Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–6 (2011).
3. Dyrløv Bendtsen, J., Nielsen, H., von Heijne, G. & Brunak, S. Improved Prediction of Signal Peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
4. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
5. Sauer, C. *et al.* Exploring the Nonconserved Sequence Space of Synthetic Expression Modules in Bacillus subtilis. *ACS Synth. Biol.* **7**, 1773–1784 (2018).
6. Macaluso, A. & Mettus, A. M. Efficient transformation of Bacillus thuringiensis requires nonmethylated plasmid DNA. *J. Bacteriol.* **173**, 1353–1356 (1991).
7. Kawamura, F. & Doi, R. H. Construction of a Bacillus subtilis double mutant deficient in extracellular alkaline and neutral proteases. *J. Bacteriol.* **160**, 442–444 (1984).
8. Lam, K. H. E., Chow, K. C. & Wong, W. K. R. Construction of an efficient Bacillus subtilis system for extracellular production of heterologous proteins. *J. Biotechnol.* **63**, 167–177 (1998).
9. Hambraeus, G., von Wachenfeldt, C. & Hederstedt, L. Genome-wide survey of mRNA half-lives in Bacillus subtilis identifies extremely stable mRNAs. *Mol. Genet. Genomics* **269**, 706–14 (2003).
10. McKenzie, T., Hoshino, T., Tanaka, T. & Sueoka, N. The nucleotide sequence of pUB110: Some salient features in relation to replication and its regulation. *Plasmid* **15**, 93–103 (1986).
11. Carl, U. D., Batz, L., Schuchardt, I., Germeroth, L. & Schmidt, T. G. M. StarGate®: A High-Capacity Expression Cloning System to Speed-Up Biopharmaceutical Development. in *Modern Biopharmaceuticals: Recent Success Stories* 147–164 (Wiley-VCH Verlag GmbH & Co. KGaA, 2013). doi:10.1002/9783527669417.ch8.
12. Femmer, C., Bechtold, M., Held, M. & Panke, S. In vivo directed enzyme evolution in nanoliter reactors with antimetabolite selection. *Metab. Eng.* **59**, 15–23 (2020).
13. Andrews, S. FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).
14. Gaspar, J. M. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19**, 536 (2018).
15. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
16. Gordon, A., Hannon, G. J. & Gordon. FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit (2014).
17. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10 (2011).
18. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. *Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States)* https://sourceforge.net/projects/bbmap/ (2014).
19. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–6 (2007).
20. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* (2012).
21. Kennard, R. W. & Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **11**, 137–148 (1969).
22. Hiromasa, K. Python implementation of Kennard-Stone algorithm. https://github.com/hkaneko1985/kennardstonealgorithm.

23. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
24. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
25. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
26. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
27. Lundberg, S. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* (2017).
28. Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **9**, 10–20 (2007).
29. Millman, K. J. & Aivazis, M. Python for scientists and engineers. *Computing in Science and Engineering* vol. 13 9–12 (2011).
30. Briggs, D. E. Gel-diffiusion method for the assay of α-amylase. *J. Inst. Brew.* **68**, 27–32 (1962).

CHAPTER 6

PROPEPTIDES: FROM PROCESSING TO PROFITING FOR PROTEIN PRODUCTION

Stefano Grasso, Liselott Schilling, Sven Hutjens,
Minia Antelo-Varela, Andreas Otto, Dörte Becher,
Tjeerd van Rij, Jan Maarten van Dijl

Abstract

One of the most common post-translational modifications occurring in Bacteria is proteolytic cleavage. Particularly, in secretory proteins this cleavage is observed for specific protein segments, named pro-peptides, which are located between the signal peptide and the mature protein. Pro-peptides are known to serve two main functions, namely a chaperone function in protein folding and an enzyme-inhibiting function. To date, the contribution of pro-peptides to protein secretion is not fully understood. Moreover, it is often not known which proteases are responsible for their cleavage. Therefore, the present study was aimed at assessing possible roles of the pro-peptides of the Bpr, Vpr and WprA proteases of *B. subtilis* in protein secretion. To this end, the signal peptides of these proteases were fused to the alkaline phosphatase PhoA of *E. coli* with or without the cognate pro-sequences. Subsequently, secretion of the PhoA protein and its activity were assessed. Furthermore, an approach was established to identify pro-peptide cleavage sites, using protease-proficient and -deficient *B. subtilis* strains in combination with mass spectrometry. Altogether, the obtained results imply that the contributions of pro-peptides to protein secretion are most likely related to a combination of specific sequence features and their functions in protein folding and enzyme activity.

Supplementary files avaialble at: https://github.com/grassoste/Thesis-supplementary-files

Introduction

In all domains of life, a significant number of the newly synthesized proteins will undergo one or more steps of post-translational modification (PTM), with cleavage and removal of a portion of the peptide being relatively common in Bacteria[1]. Cleavage of a protein may serve different purposes and, until complete processing has occurred, the protein is not considered mature. Cleavage of a protein's N- or C-terminus is particularly common amongst secreted proteins, where the best known removed sequence is the N-terminal signal peptide (SP), also known as the pre-region[2]. These SPs have a localization function by directing the respective proteins towards the pre-protein translocation machinery in the membrane. In addition, there is another type of commonly cleft sequences with more structural and enzyme inhibitory functions, namely the pro-peptide (Pro). In contrast to SPs, Pros can either be located at the N- or C-terminus of a protein, and they are usually removed once secretion is completed, i.e. in the extracellular milieu[2,3].

In some classes of secreted enzymes, e.g. proteases, both the SP and the Pro are present. In such cases, it has been shown that the Pro mainly keeps the enzyme inactive, functioning as an inhibitor until the enzyme has reached its target destination. Thus, an extracellular peptidase will not process or degrade cytosolic proteins. Also, Pros can help the cognate enzyme to fold correctly, working as an intramolecular chaperone[3–5]. For these activities, it has been demonstrated that particular Pros can act both *in cis*, i.e. when being part of the same amino acid chain, and *in trans*, i.e. as a separate molecule[6,7]. To efficiently perform such chaperone and inhibitor functions, Pros have sequence and structural features that are specific for each class of enzymes.

Pros can have interesting biotechnological applications. For instance, they can be exploited to modulate enzyme activity, stability or specificity. Such results can be obtained by means of mutagenesis or the swapping of Pros between different enzymes belonging to the same class[7]. Another possible application of Pros concerns their potential role as secretion enhancers. In fact, it has been reported multiple times that specific Pro sequences placed in-between the SP and the mature protein may increase the secretion levels, both in homologous and heterologous host organisms[8–14]. While there is evidence that Pros can somehow improve secretion, especially in *Lactococcus lactis*[10,13], there is other evidence showing that such improvements are not provided directly by the Pro itself[10]. It is in fact quite likely that the charge or hydrophobicity, or some other physico-chemical property of Pros is responsible for the improved secretion[10]. Similar to SPs[15,16], there are probably various different features, still to be elucidated, that affect secretion, since different combinations of SPs, Pros, and mature proteins yielded different results in an unpredictable manner[5,9,10,14]. One possible explanation for these observations is that the N-terminal region of the mature protein (i.e. the region after the SP cleavage site and confusingly referred to as pro-region) which encompasses up to 15-30 residues, plays an important role in modulating a protein's secretion levels[17,18]. Unfortunately, no systematic investigation has been performed so far to determine the possible mechanisms exploited by Pros to increase secretion levels, nor to clarify the role of the 'pro-region' in secretion.

The present study was aimed at elucidating the contribution of three different Pros to heterologous protein secretion in the Gram-positive bacterial cell factory *Bacillus subtilis.* To this end, the Pros were fused to the alkaline phosphatase PhoA of *Escherichia coli*, and the secretion levels of this reporter protein were subsequently evaluated. In addition, the cleavage sites between the Pros and PhoA were investigated by mass spectrometry (MS).

Material and methods

Plasmids, strains, growth media and conditions

The plasmids and strains employed in this study are listed in Table 1. *E. coli* and *B. subtilis* were grown in Lysogeny Broth (LB; 10 g/L tryptone, 10 g/L yeast extract, 10 g/L NaCl) at 37 °C and under vigorous shaking

(250 rpm). When necessary, the medium was supplemented with chloramphenicol to a final concentration of 10 µg/mL. LB plates contained 20 g/L of agar.

Plasmid and strain construction

DNA manipulations, such as DNA extraction and purification, restriction, ligation and agarose gel electrophoresis were performed with standard techniques as outlined elsewhere[23]. Cloning was performed either by overlap extension (OE)-PCR as previously described[24], or by NEB Gibson Assembly following the manufacturer's instruction. The final products were not purified, but directly used to transform *E. coli.* Competent *E. coli* cells were transformed by a 45-s heat shock at 42°C, as described elsewhere[23]. Upon successful transformation, plasmids were purified with the NucleoSpin Plasmid Mini kit (Macherey-Nagel), and the constructs were verified by PCR or sequencing. Purified plasmid DNA was used to transform *B. subtilis*, either via growth in

Strain or plasmid	Relevant properties	Reference or source
B. subtilis 168	*trpC2*	19
B. subtilis BRB08	Δ*nprB,* Δ*aprE,* Δ*epr,* Δ*bpr,* Δ*nprE,* Δ*mpr,* Δ*vpr,* Δ*wprA*	20
E. coli DH5α	*fhuA2* Δ*(argF-lacZ)U169 phoA glnV44 Φ80* Δ*(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*	21
pPSPhoA5	pLipPS2 derivative; encodes the SP-Pro sequence of a *Staphylococcus hyicus* lipase fused to the mature PhoA coding sequence of *E. coli*; Cm^r	12
pHB201	*B. subtilis - E. coli* shuttle vector carrying the P59 promoter and *cat86*::*lacZα* gene fusion; Cm^r, Em^r	22
pHB201-phoA	pHB201 derivative; encodes the only the mature *phoA* sequence from pPSPhoA5; Cm^r	This study
pHB201-bpr-phoA1	pHB201-phoA derivative; encodes a fusion between the SP-Pro of *B. subtilis* Bpr and PhoA; Cm^r	This study
pHB201-bpr-phoA2	pHB201-phoA derivative; encodes a fusion between the SP of *B. subtilis* Bpr and PhoA; Cm^r	This study
pHB201-bpr-phoA3	pHB201-phoA derivative; encodes a fusion between the Pro of *B. subtilis* Bpr and PhoA; Cm^r	This study
pHB201-bpr-SPPro	pHB201-phoA derivative; encodes the SP-Pro of *B. subtilis* Bpr; Cm^r	This study
pHB201-vpr-phoA1	pHB201-phoA derivative; encodes a fusion between the SP-Pro of *B. subtilis* Vpr and PhoA; Cm^r	This study
pHB201-vpr-phoA2	pHB201-phoA derivative; encodes a fusion between the SP of *B. subtilis* Vpr and PhoA; Cm^r	This study
pHB201-vpr-phoA3	pHB201-phoA derivative; encodes a fusion between the Pro of *B. subtilis* Vpr and PhoA; Cm^r	This study
pHB201-vpr-SPPro	pHB201-phoA derivative; encodes the SP-Pro of *B. subtilis* Vpr; Cm^r	This study
pHB201-wprA-phoA1	pHB201-phoA derivative; encodes a fusion of the SP-Pro of *B. subtilis* WprA and PhoA; Cm^r	This study
pHB201-wprA-phoA2	pHB201-phoA derivative; encodes a fusion between the SP of *B. subtilis* WprA and PhoA; Cm^r	This study
pHB201-wprA-phoA3	pHB201-phoA derivative; encodes a fusion between the Pro of *B. subtilis* WprA and PhoA; Cm^r	This study
pHB201-wprA-SPPro	pHB201-phoA derivative; encodes the SP-Pro of *B. subtilis* WprA; Cm^r	This study

Table 1. Strains and plasmids used.

Paris Medium (PM)[25] or via a modified Spizizen protocol[26].

The *E. coli phoA* gene was amplified from plasmid pPSPhoA5 with the SMF and SMR primers (all primers used in the study are reported in Supplementary Table 1). Plasmid pHB201-phoA was generated by OE-PCR using primers DMF and DMR. SP-Pro sequences were amplified with the Sx[R/F] series of primers (x can be 'B' for *bpr*, 'V' for *vpr*, or 'W' for *wprA*; R/F denotes forward and reverse primers) from the *B. subtilis* genome. From pHB201-phoA the remaining plasmids with different SPs and/or Pros were generated using the Dxx[B/V/W][F/R] primers (where xx stands either 'S' for SP only, 'P' for Pro only, or 'SP' for SP-Pro, and the remaining notation as aforementioned).

Fractionation and protein precipitation

B. subtilis was grown for 4 h in 2 mL of LB, supplemented with 10 µg/mL chloramphenicol. Subsequently, the culture was diluted 1:20 with LB to a final volume of 20 mL and incubated overnight (37 °C, 250 rpm), either in the absence or presence of mini cOmplete EDTA-free protease inhibitors (Roche). In the latter case, two tablets of the protease inhibitors were dissolved in the growth medium prior to dilution and subsequent overnight incubation. To normalize protein content over the biomass, samples equivalent to an optical density at 600 nm (OD_{600}) of 2 were collected and centrifuged at 18,620 x g for 5 min at 4 °C. The cell pellet was discarded and proteins in the supernatant were precipitated by adding trichloroacetic acid (TCA) to a final concentration of 10%. After 1 h incubation on ice, the sample was centrifuged at 18,620 x g for 10 min at 4 °C, and the supernatant was discarded. The pellet was washed with ice-cold acetone, and further centrifuged at 18,620 x g for 5 min at 4 °C. After discarding the supernatant, the pellet was dried at 60 °C for 10 min. The pellet was then resuspended in 100 µl of 1x NuPAGE LDS Sample buffer and 1x NuPAGE Sample Reducing Agent (ThermoFisher Scientific), and incubated at 95 °C for 10 min. Samples were then stored at -20 °C until further processing.

LDS-PAGE and Western blotting

Samples containing precipitated proteins from the growth medium were analyzed by LDS-PAGE, using pre-cast Bis-Tris 10% NuPAGE gels (ThermoFisher scientific). To detect all separated proteins, the gels were stained with SimplyBlue SafeStain (ThermoFisher scientific) according to the manufacturer's instructions. To detect only the PhoA constructs, unstained LDS-PAGE gels were used to semi-dry blot proteins to a 0.45 µm Amersham Protran nitrocellulose membrane (Cytiva). The membranes were then incubated overnight at 4 °C in 5% (w/v) skim milk to block aspecific binding sites. Prior to antibody incubation, the membranes were washed twice with phosphate-buffered saline supplemented with 1% Tween 20 (PBST). Subsequently, the membranes were incubated for 1 h with specific anti-PhoA polyclonal antibodies (1:5000) raised in rabbits (Eurogentec). Membranes were then washed twice in PBST, and incubated for 45 min with IRDye 800CW goat anti-rabbit IgG secondary antibodies (1:5000, LI-COR Biosciences). Finally, bound fluorescent secondary antibodies were detected with an Odyssey Infrared Imaging System (LI-COR Biosciences).

GeLC–MS analysis

To determine the cleavage sites in different constructs, samples were analyzed by GeLC-MS/MS as described previously[27]. Briefly, Coomassie-stained gels were washed in water to remove the staining. The gel was cut according to lane and protein size, and each fragment was further washed three times by incubating it at 37 °C for 15 min submerged in 750 µL of washing buffer (0.2 M ammonium bicarbonate in 30% [v/v] acetonitrile). The resulting gel fragments were dried at 30 °C in a vacuum centrifuge and subsequently rehydrated for 15 min with trypsin solution (2 µg of sequencing grade modified trypsin [Promega] in 1 mL of water). The exceeding volume of trypsin solution was discarded, and tryptic digestion was carried on at 37 °C overnight. Samples were resuspended in 100 µL of MS water and subjected to ultrasonication for 15 min to elute the

peptides from the gel pieces. The resulting supernatant was collected and transferred to an MS-vial in which samples were concentrated using a vacuum centrifuge. Once the samples were completely dry, 10 µL of Buffer A (0.1% [v/v] acetic acid) were added to the MS-vials and further used for LC-MS analysis. In-house self-packed columns were prepared as previously described[28], and used with an EASY-nLC II system (ThermoFisher Scientific). The samples were loaded onto the column with 10 µL of buffer A (0.1% [v/v] acetic acid) under a flow rate of 500 nL/min without trapping. The samples were then eluted for 100 min using a non-linear gradient with increasing concentration (from 1 to 99%) of buffer B (0.1% [v/v] acetic acid in acetonitrile) at a flow rate of 300 nL/min, and injected into the mass spectrometer. MS and MS/MS data were acquired with a Linear Trap Quadrupole Orbitrap (ThermoFisher Scientific). For data processing, the raw data was imported into MaxQuant (1.6.3.3)[29] equipped with an Andromeda[30] search engine. Database searches were carried out against a reversed *B. subtilis* 168 database[31] with manually added fusion constructs generated in the present study, and with common contaminants added by MaxQuant. Database searches were performed both against tryptic and semi-tryptic peptides. A peptides.txt file was used for further assessment of cleavage sites of the generated constructs. Protein data was exported from Scaffold and curated in Microsoft Excel before further analyses. The raw data are presented in Supplementary Table S2.

Alkaline phosphatase enzymatic assay

To detect alkaline phosphatase activity in growth medium fractions of different *B. subtilis* strains, cells were grown as described above but, after centrifugation and discarding the cell pellet, they were immediately employed in the assay. The assay was performed as previously described[32,33] with minor modifications. Briefly, 6 µL of sample were thoroughly mixed with 144 µL of a freshly prepared substrate solution (3.73 mM p-nitrophenyl phosphate [pNPP], 0.33 M diethanolamine, and 0.16 mM magnesium chloride, pH 9.8). Alkaline phosphatase activity was determined kinetically in triplicate by measuring the increase in optical density at 405 nm for 30 min at 35 s intervals, with incubation at 37 °C and under constant shaking, using a Synergy 2 Multi-Detection microplate reader (BioTek Instruments). Alkaline phosphatase from calf intestine (Sigma-Aldrich) was used to draw a set of (kinetic) standard curves for comparison to the growth medium samples. Lastly, samples from the growth medium were normalized for the biomass.

Results and discussion

Contributions of SPs and Pros to protein secretion levels

Previous studies have shown that Pros can increase the secretion levels of engineered proteins both in homologous or heterologous expression systems[8–14]. To explore the possibilities for the application of Pros to improve the secretion levels of recombinant proteins, we generated a series of fusion constructs based on *E. coli* PhoA (P00634), an alkaline phosphatase. Specifically, PhoA was fused to the SPs and Pros of three extracellular proteases of *B. subtilis*, namely Bpr, Vpr and WprA. These three proteases belong to the S8 class of subtilisin-like serine proteases. Different constructs were created, combining either only the SP, only the Pro, or both the SP and Pro with the PhoA reporter protein. In addition, constructs were created that express only the Pre-Pro peptides of Bpr, Vpr or WprA. Subsequently, the different constructs were introduced into *B. subtilis* 168 and the secretion levels were analyzed by Western blotting or alkaline phosphatase activity assays. As a control, a previously characterized PhoA fusion protein containing the SP-Pro sequence of a *S. hyicus* lipase (prepro$_{lip}$-PhoA) was included in the studies[12]. Culturing was performed either with or without addition of protease inhibitors to the growth medium to determine how proteolysis impacts on SP-Pro-PhoA processing and the secretion levels of the different constructs. Of note, the used protease inhibitors lacked a chelating agent like EDTA to avoid interference with the bacterial growth. Consequently, the activity of metalloproteases was not inhibited under the tested conditions.

As shown by Western blotting, the addition of protease inhibitors to the growth medium significantly

influenced the protein sizes of the different PhoA fusion constructs (Figure 1). This is most clearly evident for the prepro$_{lip}$-PhoA construct[12], which has a very long Pro and showed a laddering pattern with bands of higher molecular weight than that of PhoA in the presence of protease inhibitors. Similarly, the other tested constructs showed a clear shift towards a higher molecular weight product when *B. subtilis* was grown in presence of protease inhibitors. Unfortunately, the Western blots revealed no striking differences for PhoA secretion mediated by the sole SP and secretion mediated by both SP and Pro. As expected, the mere presence of the Pro was not sufficient for protein secretion. Notably, both with or without the protease inhibitors, the SP-Pro sequences of WprA did not lead to effective secretion of the reporter protein PhoA. This finding may relate to the specific structure of WprA, which is processed into two separate mature peptides (known as CWBP23 and CWBP52) through a mechanism that has not been elucidated yet. Because PhoA was fused in the same manner as mature WprA is normally fused to its SP-Pro, the expectation was that it would have been possible to detect mature PhoA in the growth medium. Several possible reasons may be entertained to explain this negative result. For instance, this may be due to the lack of a catalytically active protease *in cis* that can cleave the SP-Pro$_{WprA}$ construct, as evidenced by the fact that PhoA secretion was neither observed in the presence nor the absence of protease inhibitors. Another possible scenario is that the fusion of SP-Pro$_{WprA}$ to PhoA interfered with the release of properly folded PhoA into the growth medium, thus prompting its degradation by quality control proteases, such as HtrA, HtrB or WprA itself. In this respect, it is conceivable that the authentic WprA protease did recognize its "engineered derivative version", in spite also of its putative chaperone function and degraded it[34]. However, since the roles of the two major WprA processing products are presently still poorly understood[35], it is conceivable that more complex mechanisms take place during the processing and secretion of this protease, which ultimately affect secretion of the PhoA fusion construct.

Also, the alkaline phosphatase assays performed on the growth medium fractions with the different

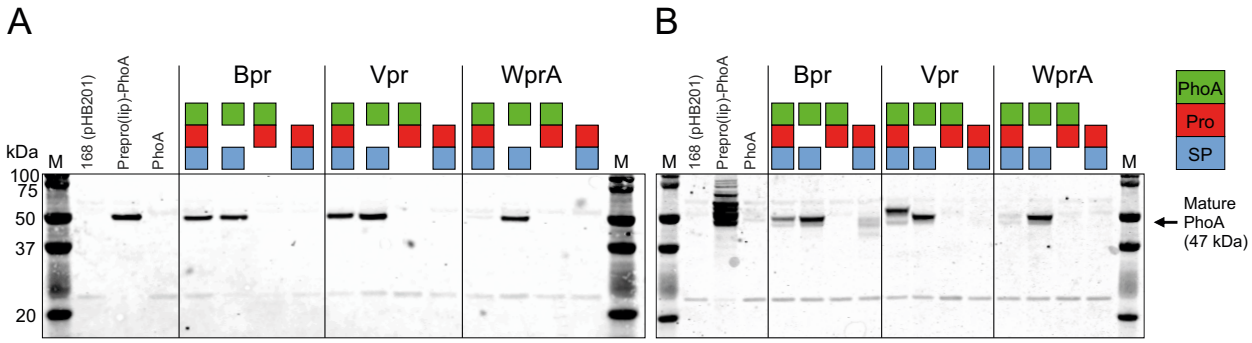


Figure 1. PhoA protein levels in the growth medium. *B. subtilis* strains expressing different combinations of Pre, Pro and/or mature PhoA were grown overnight in LB medium without **(A)** or with **(B)** protease inhibitors. Subsequently, cells were separated from the growth medium by centrifugation. Proteins in the growth medium fractions were separated by LDS-PAGE and PhoA-specific proteins were detected by Western blotting with specific polyclonal anti-PhoA antibodies. A color code is used to mark the different constructs: blue denotes the presence of a SP, red denotes the presence of a Pro, and green denotes the presence of PhoA. On top of the lanes is indicated from which of three different extracellular *B. subtilis* proteases (i.e. Bpr, Vpr or WprA) the different SP and Pro sequences were derived. On the left in panels A and B are indicated samples from control strains, including *B. subtilis* 168 harboring the empty pHB201 vector, *B. subtilis* expressing Prepro(lip)-PhoA from plasmid pPSPhoA5, and *B. subtilis* with pHB201-phoA encoding PhoA without a SP. M, molecular weight marker with the respective molecular weights indicated on the left of the gel.

secreted PhoA constructs showed different results, as presented in Figure 2. The prepro$_{lip}$-PhoA construct displayed the highest activity, in accordance with the relatively higher level of secreted PhoA and with the literature[12]. However, here one has to take into account that the prepro$_{lip}$-PhoA construct is expressed from a different plasmid and promoter. Importantly, the enzymatic assay on medium samples of bacteria grown in the presence of protease inhibitor showed that the Pros of Bpr and Vpr reduced the PhoA activity levels. This is also true for the Pro of Bpr when the bacteria were grown in absence of protease inhibitors. Conversely, the Pro of Vpr resulted in increased PhoA activity in the growth medium when the bacteria were grown in the absence of protease inhibitors. In agreement with the results from the Western blotting analysis, the SP-Pro$_{WprA}$-PhoA construct did not show any detectable PhoA activity. The differences in the PhoA activity derived from the SP-Pro$_{Bpr}$-PhoA and SP-Pro$_{Vpr}$-PhoA constructs when the bacteria were grown in absence of protease inhibitors (Figure 2) are particularly noteworthy in view of the comparable amounts of PhoA detectable by Western blotting (Figure 1). This suggests that the respective Pros may have influenced the PhoA folding process, resulting in different specific activities.

Lastly, it must be remarked that the presence of protease inhibitors led to an overall reduced PhoA activity in all samples. This may relate to lowered PhoA secretion rates in the presence of the protease inhibitors, for instance due to reduced SP degradation[34,36]. Alternatively, the protease inhibitors may influence post-translocational folding and disulfide bond formation in PhoA[37,38], or they may have an as yet undefined inhibitory effect on PhoA's enzyme activity.

Cleavage sites

To obtain additional insights into the proteases that could be involved in Pro cleavage, an MS approach was devised to identify processed products of the different SP-Pro-PhoA constructs in spent growth media of

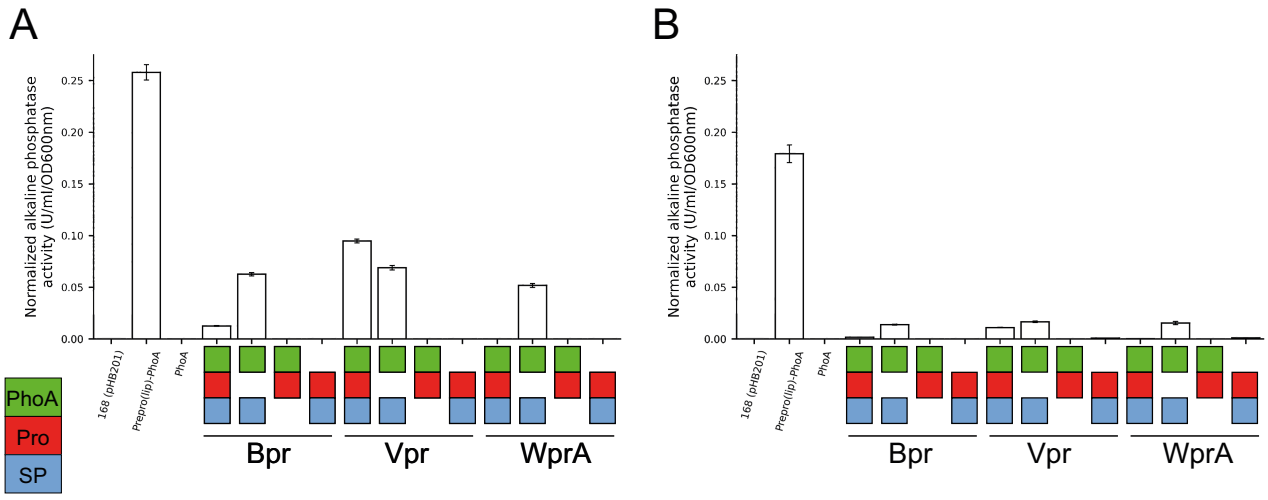


Figure 2. PhoA enzymatic activity in the growth medium. *B. subtilis* strains expressing different combinations of Pre, Pro and/or mature PhoA were grown overnight in LB medium without **(A)** or with **(B)** protease inhibitors. Subsequently, cells were separated from the growth medium by centrifugation. The growth medium fractions were used to measure alkaline phosphatase activity as shown in the bar plots. A color code is used to mark the different constructs: blue denotes the presence of a SP, red denotes the presence of a Pro, and green denotes the presence of PhoA. On top of the lanes is indicated from which of three different extracellular *B. subtilis* proteases (i.e. Bpr, Vpr or WprA) the different SP and Pro sequences were derived. On the left in panels A and B are indicated samples from control strains, including *B. subtilis* 168 harboring the empty pHB201 vector, *B. subtilis* expressing Prepro(lip)-PhoA from plasmid pPSPhoA5, and *B. subtilis* with pHB201-phoA encoding PhoA without a SP.

the *B. subtilis* reference strains 168 and BRB8. The respective strains were not only grown in the absence of protease inhibitors, but also in the presence of these inhibitors to increase the chances to detect the Pro-PhoA fusion proteins. Of note, the BRB8 strain lacks the genes for eight extracellular proteases (i.e. NprB, AprE, Epr, Bpr, NprE, Mpr, Vpr, WprA) that belong to different classes, namely M4, S8 and S1 as summarized in Table 2. If one of the eight extracellular proteases that are absent from the BRB8 strain would be responsible for Pro-PhoA cleavage, then the non-cleaved form of the respective fusion constructs should be detectable among the full tryptic fragments, possibly being the only or most abundant form. Conversely, upon Pro-PhoA cleavage, a semi-tryptic fragment beginning with the first amino acid residue after the cleavage site would be detectable. Indeed, in the medium of *B. subtilis* 168, it was possible to detect peptides corresponding with the correct cleavage sites for all eight extracellular proteases, and these were not detectable in the medium of the BRB08 strain. This demonstrated that the theoretical considerations and the experimental set-up were correct. Unfortunately, transforming the BRB8 strain with plasmids encoding the different SP-Pro-PhoA constructs of interest proved more difficult than expected. Nevertheless, in growth medium samples of the BRB08 strain, it was possible to detect full tryptic fragments spanning the cleavage site between Pro_{Vpr} and PhoA (Table 3). Surprisingly, although the peptide was detected, a spectral count of 0 was reported. Most likely, this is due to the low abundance of the fragment, as well as the simultaneous introduction of multiple fragments in the second MS. While this probably led to detection of the fragment in the second MS step, MaxQuant was unable to properly match the fragment with a spectrum from the first MS. In any case, these findings support the idea that the protease responsible for separating Pro_{Vpr} and PhoA is one of the eight that are absent from the BRB08 strain. Because of the present experimental design, it is not possible to attribute the cleavage to a specific protease, and only membrane proteases like HtrA, HtrB and RasP can be ruled out.

Of particular interest is the Pro of WprA. In fact, tryptic fragments of its CWBP23 portion were detected, which ranges approximately from position 80 to 150 in the complete amino acid sequence of WprA (Table 3). Even though this information does not provide insights into the processing sites within WprA, it shows that in *B. subtilis* 168 CWBP23 is not degraded, regardless of the presence or absence of protease inhibitors[39]. Unfortunately, it is presently neither possible to precisely attribute this region to the wild-type WprA nor to the SP-Pro_{WprA}-PhoA construct. In fact, fragments from the PhoA part of SP-Pro_{WprA}-PhoA were detected by MS. This could possibly mean that CWBP23 has its own function[35], independent from the protease domain of WprA (CWBP52). Of note, peptides from the connecting region between, approximately, residues 160 to 550 were never retrieved, suggesting that this could be the actual Pro-region of CWBP52, which is degraded upon cleavage.

Protease	Family
NprB	M4
AprE	S8
Epr	S8
Bpr	S8
NprE	M4
Mpr	S1
Vpr	S8
WprA	S8

Table 2. Proteases missing from B. subtilis BRB08 and the respective protease families.

Conclusion and future perspectives

With the present approach, it was possible to shed some more light on both the industrial applicability of Pros, as well as their basic functioning. Primarily, the present findings are in agreement with a relevant body of literature[2,5,9,10,14,17,18] that claims that Pros have roles in protein folding and enzyme inhibition, and that Pros merely influence secretion because they are the *de facto* pro-region of the protein during the secretion process. In other words, the combined Pro-mature protein represents the polypeptide that undergoes membrane translocation, post-translocational folding and release into the growth medium. Thus, previous claims that Pros can be used to enhance secretion[8–14]

Description	Sample	Sequence	Amino acid before*	Length	Missed cleavages	Mass	Leading razor protein	Start position	End position	PEP	Score	Intensity	MS/MS Count
Fragment on the cl. Site of SP-Pro$_{Vpr}$-PhoA	BRB+PI	TDNMKDKDVTISEDAVSPQMDDSMYNR	K	27	2	3104.34	VP	142	168	0.00	50.27	0	0
Fragment on the cl. Site of SP-Pro$_{Vpr}$-PhoA	168+PI	DKDVTISEDAVSPQMDDSMYNR	K	22	1	2515.08	VP	147	168	0.00	124.94	0	0
Fragment on the cl. Site of SP-Pro$_{Vpr}$-PhoA	168+noPI	DKDVTISEDAVSPQMDDSMYNR	K	22	1	2515.08	VP	147	168	0.0	170.22	0	0
Fragment on the cl. Site of SP-Pro$_{Vpr}$-PhoA	BRB+PI	DKDVTISEDAVSPQMDDSMYNR	K	22	1	2515.08	VP	147	168	0.00	160.05	0	0
Fragment on the cl. Site of SP-Pro$_{Vpr}$-PhoA	168+PI	DKDVTISEDAVSPQMDDSMYNR	K	22	1	2515.08	VP	147	168	0.00	124.94	0	0
Fragment on the cl. Site of SP-Pro$_{Vpr}$-PhoA	168+noPI	DKDVTISEDAVSPQMDDSMYNR	K	22	1	2515.08	VP	147	168	0.0	170.22	0	0
Fragment on the cl. Site of SP-Pro$_{Vpr}$-PhoA	BRB+PI	DKDVTISEDAVSPQMDDSMYNR	K	22	1	2515.08	VP	147	168	0.0	160.05	0	0
Fragment on the cl. Site of SP-Pro$_{Vpr}$-PhoA	BRB+noPI	DKDVTISEDAVSPQMDDSMYNR	K	22	1	2515.08	VP	147	168	0.01	48.90	0	0
Fragment from CWBP23 (based on predicted MW)	168+PI	PGATDIQK	T	8	0	828.43	WP	64	71	0.01	148.04	5721200	2
Fragment from CWBP23 (based on predicted MW)	168+noPI	SDSVLNVSVYPSK	K	13	0	1393.71	WP	81	93	0.00	134.56	22732000	3
Fragment from CWBP23 (based on predicted MW)	168+PI	SDSVLNVSVYPSK	K	13	0	1393.71	WP	81	93	0.00	140.91	20929000	3
Fragment from CWBP23 (based on predicted MW)	168+PI	SDSVLNVSVYPSKEK	K	15	1	1650.85	WP	81	95	0.00	133.05	2411200	1
Fragment from CWBP23 (based on predicted MW)	168+PI	SDSVLNVSVYPSK	K	13	0	1393.71	WP	81	93	0.00	140.91	21116000	6
Fragment from CWBP23 (based on predicted MW)	168+PI	SDSVLNVSVYPSKEK	K	15	1	1650.85	WP	81	95	0.00	133.05	11972000	2
Fragment from CWBP23 (based on predicted MW)	168+noPI	SDSVLNVSVYPSK	K	13	0	1393.71	WP	81	93	0.00	134.56	23361000	7

(continued)

Fragment from CWBP23 (based on predicted MW)	168+noPI	SDSVLNVSVYPSKEK	K	15	1	1650.85	WP	81	95	0.04	38.67	4477500	1
Fragment from CWBP23 (based on predicted MW)	168+PI	SDSVLNVSVYPSK	K	13	0	1393.71	WP	81	93	0.00	165.58	27127000	3
Fragment from CWBP23 (based on predicted MW)	168+noPI	ALKDETFEMYR	K	11	1	1401.66	WP	96	106	0.00	146.71	0	0
Fragment from CWBP23 (based on predicted MW)	168+PI	ALKDETFEMYR	K	11	1	1401.66	WP	96	106	0.01	118.03	6378400	1
Fragment from CWBP23 (based on predicted MW)	168+PI	ALKDETFEMYR	K	11	1	1401.66	WP	96	106	0.00	118.03	6378400	1
Fragment from CWBP23 (based on predicted MW)	168+noPI	ALKDETFEMYR	K	11	1	1401.66	WP	96	106	0.00	146.71	5360500	1
Fragment from CWBP23 (based on predicted MW)	168+PI	ALKDETFEMYR	K	11	1	1401.66	WP	96	106	0.01	118.28	0	0
Fragment from CWBP23 (based on predicted MW)	168+noPI	VEYLGEEEPEDGGTAEAAAEAK	K	22	0	2264.00	WP	131	152	0.0	276.59	15687000	6
Fragment from CWBP23 (based on predicted MW)	168+PI	VEYLGEEEPEDGGTAE	K	16	0	1722.71	WP	131	146	0.10	61.21	335630	0
Fragment from CWBP23 (based on predicted MW)	168+PI	VEYLGEEEPEDGGTAEAAAEAK	K	22	0	2264.00	WP	131	152	0.0	266.28	18409000	3
Fragment from CWBP23 (based on predicted MW)	168+PI	VEYLGEEEPEDGGTAEAAAEAK	K	22	0	2264.00	WP	131	152	0.0	266.28	38638000	6
Fragment from CWBP23 (based on predicted MW)	168+noPI	VEYLGEEEPEDGGTAEAAAEAK	K	22	0	2264.00	WP	131	152	0.0	276.59	21035000	8
Fragment from CWBP23 (based on predicted MW)	168+PI	VEYLGEEEPEDGGTAE	K	16	0	1722.71	WP	131	146	0.02	103.31	690110	1
Fragment from CWBP23 (based on predicted MW)	168+PI	VEYLGEEEPEDGGTAEAAAEAK	K	22	0	2264.00	WP	131	152	0.0	264.85	26873000	8
Fragment from CWBP23 (based on predicted MW)	BRB+PI	SIRDEQLSQTAEGK	R	14	1	1560.77	WP	194	207	0.00	195.72	23751000	6
Fragment from CWBP23 (based on predicted MW)	BRB+PI	SIRDEQLSQTAEGK	R	14	1	1560.77	WP	194	207	0.0	195.72	39996000	11

(continued)

Fragment from CWBP23 (based on predicted MW)	BRB+PI	DEQLSQTAEGK	R	11	0	1204.56	WP	197	207	0.00	128.88	382950	2
Fragment from CWBP23 (based on predicted MW)	BRB+PI	DEQLSQTAEGK	R	11	0	1204.56	WP	197	207	0.00	128.88	382950	2
Fragment possibly inbe-twween CWBP23 and CWBP52 (based on predicted MW)	BRB+PI	TAGAILTENNIAAK	K	14	0	1385.75	WP	302	315	0.0	199.19	5454400	2
Fragment possibly inbe-twween CWBP23 and CWBP52 (based on predicted MW)	BRB+PI	TAGAILTENNIAAK	K	14	0	1385.75	WP	302	315	0.0	199.19	5667500	2

Table 3. Summary of most interesting peptides identified by GeLC-MS/MS. No PI, absence of protease inhibitors. PI, protease inhibitors were present during culturing. *Yellow = full tryptic peptides; Green = semi-tryptic peptides

may have to be reconsidered, for instance by critically analyzing the physico-chemical properties of the region immediately downstream of the SP cleavage site. Biotechnological efforts aimed at improving protein secretion rates should, therefore, include rational engineering approaches where different physico-chemical properties of the pro-region are tested. Conversely, studies on Pros should rather be aimed at modifying, tuning, and/or improving the enzyme activity of naturally occurring Pro-mature protein pairs, as already discussed in literature[5,7]. In parallel, such rational approaches would yield relevant information on the relative contributions of different features of the Pros to protein secretion and enzyme folding, in particular with respect to sequence-structure combinations and physico-chemical properties of the sequence.

The other conclusion that can be drawn from the present study is that one of the eight proteases absent from *B. subtilis* BRB08 (i.e. NprB, AprE, Epr, Bpr, NprE, Mpr, Vpr, WprA) is responsible for cleavage of at least the Pro from Vpr. Unfortunately, additional studies will be necessary to fully understand the relationship(s) occurring between proteases and Pros in *B. subtilis*. No information is currently available whether these proteases are activated in a complex cascade-like fashion, similar to aureolysin and other extracellular proteases of *Staphylococcus aureus*[40–42], or if autolysis plays a major role. Additionally, even though the actual proteases responsible for the various cleavage events are not known, they are inhibited by the employed cocktail of protease inhibitors. This implies that, most likely, metalloproteases play only a marginal role in Pro cleavage as the employed protease inhibitor cocktail lacked metalloprotease inhibitors. Of course, in view of the previously elucidated complex processing networks exemplified for *S. aureus*, it will be challenging to attribute specific activities in Pro processing to particular proteases of *B. subtilis*. So far, the *in vivo* roles of extracellular proteases of *B. subtilis* have mostly been investigated with regards to the degradation of secreted proteins, nutrient scavenging, and competence-, sporulation-, biofilm- and swarming-related functions[43–47]. Altogether, it can be concluded that, to efficiently exploit Pros in biotechnological applications for protein secretion or enzyme optimization, a clearer understanding of their general modes of action is needed. This calls for a more thorough functional dissection of the different extracellular proteases of *B. subtilis*, and other Pro-containing secretory proteins within their natural context.

References

1. Cain, J. A., Solis, N. & Cordwell, S. J. Beyond gene expression: the impact of protein post-translational modifications in bacteria. *J. Proteomics* **97**, 265–86 (2014).
2. Tjalsma, H. *et al.* Proteomics of Protein Secretion by Bacillus subtilis : Separating the 'Secrets' of the Secretome. *Microbiol. Mol. Biol. Rev. 2* **68**, 207–233 (2004).
3. Braun, P. & Tommassen, J. Function of bacterial propeptides. *Trends Microbiol.* **6**, 6–8 (1998).
4. Wandersman, C. Secretion, processing and activation of bacterial extracellular proteases. *Mol. Microbiol.* **3**, 1825–1831 (1989).
5. Demidyuk, I. V., Shubin, A. V., Gasanov, E. V. & Kostrov, S. V. Propeptides as modulators of functional activity of proteases. *Biomol. Concepts* **1**, 305–322 (2010).
6. Corvey, C., Stein, T., Düsterhus, S., Karas, M. & Entian, K. D. Activation of subtilin precursors by Bacillus subtilis extracellular serine proteases subtilisin (AprE), WprA, and Vpr. *Biochem. Biophys. Res. Commun.* **304**, 48–54 (2003).
7. Takagi, H. & Takahashi, M. A new approach for alteration of protease functions: pro-sequence engineering. *Appl. Microbiol. Biotechnol.* **63**, 1–9 (2003).
8. Le Loir, Y. *et al.* Signal Peptide and Propeptide Optimization for Heterologous Protein Secretion in Lactococcus lactis. *Appl. Environ. Microbiol.* **67**, 4119–4127 (2001).
9. Lim, P. Y., Tan, L. L., Ow, D. S. W. & Wong, F. T. A propeptide toolbox for secretion optimization of Flavobacterium meningosepticum endopeptidase in Lactococcus lactis. *Microb. Cell Fact.* **16**, 221 (2017).
10. Morello, E. *et al.* Lactococcus lactis, an efficient cell factory for recombinant protein production and secretion. *J. Mol. Microbiol. Biotechnol.* **14**, 48–58 (2007).
11. Kakeshita, H., Kageyama, Y., Ara, K., Ozaki, K. & Nakamura, K. Propeptide of Bacillus subtilis amylase enhances extracellular production of human interferon-α in Bacillus subtilis. *Appl. Microbiol. Biotechnol.* **89**, 1509–17 (2011).
12. Kouwen, T. R. H. M. *et al.* Contributions of the Pre- And Pro-Regions of a Staphylococcus hyicus Lipase to Secretion of a Heterologous Protein by Bacillus subtilis. *Appl. Environ. Microbiol.* **76**, 659–669 (2010).
13. Kakeshita, H., Kageyama, Y., Ozaki, K., Nakamura, K. & Ar, K. Improvement of Heterologous Protein Secretion by Bacillus subtilis. in *Advances in Applied Biotechnology* (InTech, 2012). doi:10.5772/29256.
14. Le Loir, Y. *et al.* Protein secretion in Lactococcus lactis : an efficient way to increase the overall heterologous protein production. *Microb. Cell Fact.* **4**, 2 (2005).
15. Brockmeier, U. *et al.* Systematic Screening of All Signal Peptides from Bacillus subtilis: A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-positive Bacteria. *J. Mol. Biol.* **362**, 393–402 (2006).
16. Degering, C. *et al.* Optimization of Protease Secretion in Bacillus subtilis and Bacillus licheniformis by Screening of Homologous and Heterologous Signal Peptides. *Appl. Environ. Microbiol.* **76**, 6370–6376 (2010).
17. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.* **97**, 422–441 (2018).
18. Tian, P. & Bernstein, H. D. Identification of a post-targeting step required for efficient cotranslational translocation of proteins across the Escherichia coli inner membrane. *J. Biol. Chem.* **284**, 11396–11404 (2009).
19. Kunst, F. *et al.* The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature* **390**, 249–256 (1997).

20. Pohl, S. *et al.* Proteomic analysis of Bacillus subtilis strains engineered for improved production of heterologous proteins. *Proteomics* **13**, 3298–3308 (2013).
21. Taylor, R. G., Walker, D. C. & McInnes, R. R. E. coli host strains significantly affect the quality of small scale plasmid DNA preparations used for sequencing. *Nucleic Acids Res.* **21**, 1677–8 (1993).
22. Bron, S. *et al.* Protein secretion and possible roles for multiple signal peptidases for precursor processing in Bacilli. *J. Biotechnol.* **64**, 3–13 (1998).
23. Russell, D. W. & Sambrook, J. *Molecular cloning: a laboratory manual*. vol. 1 (Cold Spring Harbor Laboratory Cold Spring Harbor, NY, 2001).
24. Bryksin, A. V. & Matsumura, I. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. *Biotechniques* **48**, 463–5 (2010).
25. Monteferrante, C. G., Miethke, M., Van Der Ploeg, R., Glasner, C. & Van Dijl, J. M. Specific targeting of the metallophosphoesterase YkuE to the Bacillus cell wall requires the twin-arginine translocation system. *J. Biol. Chem.* **287**, 29789–29800 (2012).
26. Sauer, C. *et al.* Exploring the Nonconserved Sequence Space of Synthetic Expression Modules in Bacillus subtilis. *ACS Synth. Biol.* **7**, 1773–1784 (2018).
27. Bonn, F. *et al.* Picking vanished proteins from the void: How to collect and ship/share extremely dilute proteins in a reproducible and highly efficient manner. *Anal. Chem.* **86**, 7421–7427 (2014).
28. Bernal-Cabas, M. *et al.* Functional association of the stress-responsive LiaH protein and the minimal TatAyCy protein translocase in Bacillus subtilis. *Biochim. Biophys. Acta - Mol. Cell Res.* **1867**, 118719 (2020).
29. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
30. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
31. Barbe, V. *et al.* From a consortium sequence to a unified sequence: The Bacillus subtilis 168 reference genome a decade later. *Microbiology* **155**, 1758–1775 (2009).
32. Mössner, E., Boll, M. & Pfleiderer, G. Purification of human and bovine alkaline phosphatases by affinity chromatography. *Hoppe. Seylers. Z. Physiol. Chem.* **361**, 543–9 (1980).
33. Bergmeyer, H. U. *Methods of enzymatic analysis*. (Verlag Chemie, 1974).
34. Krishnappa, L., Monteferrante, C. G., Neef, J., Dreisbach, A. & van Dijl, J. M. Degradation of Extracytoplasmic Catalysts for Protein Folding in Bacillus subtilis. *Appl. Environ. Microbiol.* **80**, 1463–1468 (2014).
35. Stephenson, K. & Harwood, C. R. Influence of a cell-wall-associated protease on production of alpha-amylase by Bacillus subtilis. *Appl. Environ. Microbiol.* **64**, 2875–2881 (1998).
36. Neef, J., Bongiorni, C., Goosens, V. J., Schmidt, B. & van Dijl, J. M. Intramembrane protease RasP boosts protein production in Bacillus. *Microb. Cell Fact.* **16**, 57 (2017).
37. Kouwen, T. R. H. M. *et al.* Thiol-disulphide oxidoreductase modules in the low-GC Gram-positive bacteria. *Mol. Microbiol.* **64**, 984–99 (2007).
38. Kouwen, T. R. H. M., Dubois, J.-Y. F., Freudl, R., Quax, W. J. & van Dijl, J. M. Modulation of thiol-disulfide oxidoreductases for increased production of disulfide-bond-containing proteins in Bacillus subtilis. *Appl. Environ. Microbiol.* **74**, 7536–45 (2008).
39. Antelmann, H. *et al.* A proteomic view on genome-based signal peptide predictions. *Genome Res.* **11**, 1484–1502 (2001).
40. Shaw, L., Golonka, E., Potempa, J. & Foster, S. J. The role and regulation of the extracellular proteases of Staphylococcus aureus. *Microbiology* **150**, 217–228 (2004).
41. Tam, K. & Torres, V. J. Staphylococcus aureus Secreted Toxins and Extracellular Enzymes. in *Gram-*

Positive Pathogens 640–668 (ASM Press, 2019). doi:10.1128/9781683670131.ch40.

42. Gimza, B. D., Larias, M. I., Budny, B. G. & Shaw, L. N. Mapping the Global Network of Extracellular Protease Regulation in Staphylococcus aureus. *mSphere* **4**, e00676-19 (2019).
43. Connelly, M. B., Young, G. M. & Sloma, A. Extracellular proteolytic activity plays a central role in swarming motility in Bacillus subtilis. *J. Bacteriol.* **186**, 4159–4167 (2004).
44. Krishnappa, L. *et al.* Extracytoplasmic proteases determining the cleavage and release of secreted proteins, lipoproteins, and membrane proteins in Bacillus subtilis. *J. Proteome Res.* **12**, 4101–4110 (2013).
45. Pohl, S. & Harwood, C. R. Heterologous protein secretion by bacillus species from the cradle to the grave. *Adv. Appl. Microbiol.* **73**, 1–25 (2010).
46. Sarvas, M., Harwood, C. R., Bron, S. & van Dijl, J. M. Post-translocational folding of secretory proteins in Gram-positive bacteria. *Biochim. Biophys. Acta* **1694**, 311–27 (2004).
47. Stephenson, S., Mueller, C., Jiang, M. & Perego, M. Molecular analysis of Phr peptide processing in Bacillus subtilis. *J. Bacteriol.* **185**, 4861–71 (2003).

CHAPTER 7

GENERAL DISCUSSION AND FUTURE PERSPECTIVES

General discussion

Protein secretion has been industrially exploited during the last four decades, carrying along a continuous improvement of (micro-)organisms, secretion pathways (e.g. by bottleneck removal), and overall production processes from up- to down-stream. Nevertheless, while some technical and engineering aspects have been deeply and formally investigated (e.g. fermentation techniques or biomass balances), other aspects, mostly purely biological, were often tackled via trial-and-error, but without a more general model, theory, or simply understanding behind them. This has led to the paradox of being able to exploit microorganisms, but only after selecting them through high-throughput (HT) screenings rather than a rational "construction" process. Imagine if, to build a circuit board or an engine, engineers would generate thousands of them with small differences to then test them in real word-like conditions. On the contrary, the engineers' approach is generally to ideate, design and simulate a novel circuit board or engine, and only then to prototype it and test it for production. While clearly biotechnology and engineering are based on different disciplines, i.e. biology and physics, respectively, this does not mean that models and rules cannot be devised also in the living world.

Over the last decade the idea spread that the principles of engineering can also be applied to biological systems, with the consequence that more and more biological models have been created. This has mostly been emphasized and explored for model systems, such as *Escherichia coli*[1], *Saccharomyces cerevisiae*[1] and human cells[2]. Unfortunately, other organisms that are also relevant, for instance for human health or industrial production, are often less well studied from theoretical and modelling perspectives. *Bacillus subtilis* presents an intermediate situation: because of its industrial relevance, it is well studied and heavily exploited. Nevertheless, many fundamental aspects, not directly or immediately linked to exploitation, have still been overlooked and left behind. Other organisms, such as the pathogens *Staphylococcus aureus* and *Porphyromonas gingivalis* represent important threats for human health and wellbeing, but have only been marginally investigated from a theoretical angle. Therefore, the research presented in this thesis was aimed at bridging the gap between computational and experimental approaches in studies on the industrial microorganism *B. subtilis* and the pathogens *S. aureus* and *P. gingivalis* with a major focus on protein sorting and secretion.

The traditional approach to create models relies heavily on mathematics. The big advantage of mathematical models is that they are particularly clear and transparent, which allows one to answer questions about the particular model itself. More recently, a new branch of modelling developed, which is still based on mathematics, but implemented through artificial intelligence (AI). This has resulted in popular approaches, such as machine learning (ML) or deep learning. Building such AI models has definitely many advantages. In particular, this approach does not require any knowledge of the system to be modelled, and it produces relatively accurate predictions. The disadvantage is that models of this kind require huge datasets to be used for training and testing, and that they ultimately represent 'black-boxes' that do not lead to mechanistic insights[1,3]. For such reasons, mathematical modelling has not been abandoned and still plays an important role[4–6], especially when a sufficient level of understanding and prior knowledge is available. In any case, especially within biology, there is a need for interpretable, or 'white box', models that clarify the underlying principles and mechanisms[7]. This principle was central for the studies presented in **Chapters 2** to **5** of this thesis, where interpretability was conjugated with specific experimental and theoretical approaches.

Chapters 2 to **4** present the implementation of protein subcellular localization (SCL) prediction pipelines based on tailored meta-tools that exploit different software (SW) for specific tasks. More precisely, the developed pipelines were designed following a sorting signal-based approach. This means that the amino acid sequences of proteins are searched for specific domains, motifs or fingerprints that can point toward their final SCL. The most common alternative approach used nowadays is the homology-based transfer of knowledge (i.e. the SCL), which relies on the experimental annotation in public databases (DBs) of a protein

homologous to the query protein[8,9]. The two approaches have their respective pros and cons, but they differ mainly on one important point, namely their capability to explain the final prediction. While homology-based approaches can only inform us that they identified a homologous protein with an annotated SCL in a given DB, sorting signal-based approaches provide striking causal information on the final SCL prediction by pointing out which signals are responsible for sorting the protein to a specific subcellular compartment.

A frequently employed present-day approach for protein SCL prediction is to employ ML, or similar algorithms, to the underlying features that describe a query protein. While ML proved to be a very powerful tool, the final predictions should be taken *per se*, with the only addition of a probability level, and leaving the user with nothing to interpret. Although leaving space for interpretation to the user could be seen as a lack of confidence in the prediction, it actually allows a prediction tool to be more flexible and to yield more precise results. Prediction tools, including the GP[4] described in **Chapter 2**, are usually built around one or few model species. Consequently, they are intrinsically biased towards a plessocentric (from πλάσσω, mould in Greek) view of protein sorting. Hence, the possibility to re-interpret the results by providing the necessary information should be seen as an improvement towards more tailored predictions. This is exemplified in **Chapter 3**, where a specific and tailored SCL prediction tool for the oral pathogen *P. gingivalis* is described. Prediction tools for Gram-negative bacteria in general would, in fact, be ineffective in predicting a specific and medically relevant subset of proteins that are secreted via the Por secretion system (PorSS or T9SS), which is uniquely present in *P. gingivalis* and related species, where it was exapted from a gliding motility apparatus[10]. On the contrary, results presented in **Chapter 4**, predicting a high number of cytosolic proteins within the growth medium of *S. aureus*, represent a clear example of our limited current understanding of all the secretion mechanisms employed by this pathogen. The high number of extracellular cytoplasmic proteins (ECPs) can in fact be explained in multiple alternative ways. For instance, there may still be unknown secretion pathways in *S. aureus* or secretion pathways whose substrates cannot be predicted yet. Alternatively, there are ways of secreting cytosolic proteins through cell lysis[11] or other undetected methods[12]. In any case, the more it is clarified how predictions are made and allow a critical evaluation by the user, the more biologically meaningful will be the results. Conversely, if the results from predictions are based on data and features different from the sorting signals, the biological relevance will remain unclear. In these cases, one actually has to blindly trust the SCL of the homologous protein hit in case this is explicitly specified, or the physico-chemical properties associated with the respective SCL prediction.

Information on the specific features taken into account for a particular SCL prediction can be viewed as a first layer of interpretability. However, a second layer that integrates the various detected features may hide the logic of the final SCL assignment. In fact, even if the features on which an AI prediction method is based are known, the respective importance of each feature and how to “calculate” the final results may not be known, not even to the authors. This problem is typical for ML models. With ML models it is possible to predict SCLs, even with striking results, but they do not provide any information that can be used by scientists to hypothesize or progress a theory (which is *de facto* a model). This limitation in understanding the importance of different underlying features may have implications also for determining the applicability range of a ML model. As the model is built on a dataset provided for training, any future predicted element should be similar to a component of the training set. Moreover, there is no clear boundary to indicate where the applicability of the model ceases to be reliable. Thus, clear weights, as implemented in the SCL prediction tools presented in **Chapters 2** to **4**, can drastically help in the interpretation of results and, in turn, to define or refine a specific biological theory. The approach exploited in the three respective prediction methods, namely exploiting sorting signals and simple score assignments, are likely to produce some of the most transparent and interpretable SCL predictions possible.

Unfortunately, it is not always feasible to design and implement simple models, and very often ML can help in detecting patterns, associations and other relevant information that is embedded in huge training

datasets, which human eyes and brains would not be able to detect and appreciate. This is for instance the case when there is a very high number of features that are relevant to a specific problem or model, and that need to be taken into account. **Chapter 5** presents a clear example of such a problem in relation to investigations on the role of the signal peptide (SP) sequence in the efficiency of protein secretion. Importantly, the efficiency of protein secretion directed by a SP seems not to be explainable with a single feature, both without[13–15] and with the employment of ML[16]. On the contrary, previous studies on this topic[15,16] showed: i) that secretion efficiency is also determined by properties of the RNA transcript, such as the minimum folding energy; and ii) that combining multiple features improves the predictability of secretion efficiency. Clearly, also the size of the entry data used in such studies, which was in the order of one or two hundred SPs, proved to be insufficient for attempts to solve the intricate puzzle of what determines SP efficiency.

To generate an interpretable model for SP efficiency, taking into account multiple features and based on a sufficient amount of data, the approach described in **Chapter 5** was devised. Briefly, a library of ~12,000 unique SPs was fused to the reporter amylase AmyQ, introduced into *B. subtilis* and screened in a HT fashion to determine the secretion efficiency of each SP. The resulting data was then used, together with an array of 156 physico-chemical features describing each SP both at the amino acid and nucleotide levels, to generate a model that explains the most relevant characteristics of a SP and how to exploit them to improve the efficiency of protein secretion. With little quantitative prior data on protein secretion efficiency and the high dimensionality of the problem, a mathematical model would have proven difficult to devise. Therefore, a solution was sought in combining ML with an interpretation analysis. Importantly, with an interpretation analysis it is possible to study the 'black-box' that constitutes a ML model, and to understand how the model 'transforms' the input data into the output of the model. To exemplify, the model presented in **Chapter 5** addresses the 156 physico-chemical features describing any SP in relation to its efficiency in directing protein secretion. Today, model interpretation is still quite challenging[17], and only few tools are available. One of these tools is SHAP[18–20]. By implementing a SHAP analysis, the afore-mentioned points were successfully tackled through the generaton of a model that can predict protein secretion efficiency, but also provides explanations of the features that make up an efficient SP. Specifically, the model explains for each individual SP the contribution of each physico-chemical feature to its efficiency in directing secretion. Due to the global and local explainability provided by SHAP, the model can be exploited both for basic science to explain the theory behind SP efficiency, and for applied purposes to predict, provide insights and in turn tweak the secretion efficiency of a specific amino acid sequence to be used as SP. While the combined usage of ML and its interpretation may not be completely flawless, for instance not being able to distinguish between real causality and model-driven artifacts, it is still one of the best applicable options[21]. This is especially relevant for addressing biological problems, where there is no other possibility to develop a model, e.g. through purely mathematical approaches.

Among the limitations of the studies documented in **Chapters 2** to **5**, the most relevant one relates to the availability and fairness of data or knowledge that is used to train or build the model[22,23], and consequently its applicability space[24]. As discussed in **Chapters 2** and **3**, the numbers of proteins with known SCL in public DBs is limited and, often, the annotation is flawed either because information was incorrectly reported, or as a consequence of poor experimental data. This latter issue is actually very relevant and should be regarded as a limitation of the currently available experimental techniques. To exemplify, if one were to take the data from the most comprehensive proteome fractionation study in *B. subtilis*[25], one would find that the majority of the identified proteins (52% of the proteome) was detected in multiple fractions, with a non-negligible number of proteins detected even in 4 or 5 different fractions. While this is not entirely unexpected, since proteins are dynamic entities that 'travel' through the cell to reach their ultimate SCL, the data could not be used to successfully train a ML model. Optionally, it would be possible to clean and refine this dataset, but the inherent risk would be to introduce biases. Additionally, questions would arise on its potential applicability,

especially if the non-identified proteins would create a defined cluster. For instance, supposed that all the non-identified proteins were proteases, the main consequence would be that the model is not applicable to predict the SCL of proteases, and thus the model could not be trusted in such an 'area' of the predictions. In addition to interpretability, the limited number of training data has been another reason why the current SCL predictors were all devised as expert system predictors, rather than ML models. Nevertheless, while the different implemented approaches make the precise evaluation of the prediction accuracy harder than in a ML model, this does not mean that they can be safely employed beyond the respective boundaries of applicability.

The afore-mentioned issues also apply to the model for SP efficiency as presented in **Chapter 5**. While in this case a significant amount of data was used to train the model, its size still remains suboptimal. Additionally, the design space was not perfectly uniformly sampled, meaning that not all the possible values and their combinations within the design space (i.e. the 156 physico-chemical features) were sampled. Also, those combinations that were sampled may have different degrees of representativity within the dataset. Luckily, the higher the number of data used, as well as the numbers of measurements, the more these effects will be averaged out. In fact, the presented model proves to be accurate both when tested *in silico* and *in vivo*, while the residual error can be ascribed to a mix of biases in the design of the SP library and the experimental setup. While higher numbers of data points and improved designs can effectively reduce biases and increase predictability, they cannot extend a model beyond its boundaries[24]. For instance, the studies in **Chapter 5** were designed and performed in *B. subtilis* grown within a nanoliter reactor (NLR) system. Consequently, the findings regarding SP-directed protein secretion efficiency cannot be directly transferred to different organisms or growth conditions. To achieve this, it would be necessary either to experimentally validate the model under different conditions, or at least to know the relationships between the model conditions and the conditions of interest, in order to make corrections or to determine conversion factors. Unfortunately, there is currently insufficient quantitative knowledge about SP efficiency under different experimental conditions, leaving as sole option the experimental validation.

While the mentioned limitations in terms of applicability, experimental accuracy, dataset size may seem overwhelming, the overall approach described by the design-build-test-learn (DBTL) cycle[26] yields the best approximation of reality. This DBTL cycle, as exemplified by the first round described in **Chapter 5**, is mainly exploited in the fields of synthetic biology, metabolic engineering and strain improvement, mostly with an applied industrial objective. In such cases, especially when ML methods are adopted within the cycle, the 'learning' step belongs to the machine (i.e. the ML model) and not to the human operator, causing an actual loss of knowledge and understanding. Differently, by interpreting the ML model, or by exploiting a 'white-box' model, the outcome offers two options. Firstly, the model can immediately be exploited for the subsequent cycle or for other predictive purposes while, alternatively, its interpretation can yield interesting insights into the underlying biological mechanisms. To date, this approach is mostly applied in biological engineering, but it would benefit basic science if interpreted or interpretable models would become open access. In such a way, multiple models characterizing, and thus potentially explaining, different biological components of a system, such as promoters, repressors, switches, SPs, domains and catalytic sites, could be taken into consideration in an integrated manner. This would improve our understanding of the bigger picture through a model comprising multiple components.

Major limiting factors that slow down the creation of models are the availability and reliability of data. As mentioned above, there is a need for high amounts of data. Even though the studies described in **Chapters 2** to **4** did not employ ML in the respective models, a higher number of trustworthy properly annotated proteins would have been beneficial in both the design and testing of these models. This demand for big numbers is even higher when ML models are employed. However, also the fairness of data becomes important within this context[22,23]. For such reasons, a good model design is important to avoid possible biases,

and standardisation of the design and experimental set up will help in comparing or merging multiple datasets. Obviously, it is not always trivial to practically implement these principles within an experimental set up, also due to technical limitations or other logistic constraints. An example of this was already discussed in regard to the SCL of *B. subtilis* proteins[25]. However, with respect to the HT screening of protein secretion, dramatic advances have been made since 2010[14], especially with the advent of microfluidics and other technologies, such as the NLRs described in **Chapter 5**.

The view that the mentioned limitations in predicting SCLs or SP efficiency can be overcome, is underpinned by the significance and robustness of the approaches presented in this thesis. For example, the GP[4] prediction tool presented in **Chapter 2** shows a striking degree of accuracy when benchmarked over a test dataset of experimentally determined SCLs of proteins from many species belonging to two different *phyla*, namely the Actinobacteria and Firmicutes. Unfortunately, for the results presented in **Chapter 3**, a thorough validation on an external dataset was not possible. Nonetheless, the study demonstrates how the SCL prediction, together with a complete functional analysis, is a powerful tool to detect determinants for bacterial fitness and virulence. The results will thus be useful for clinical and biochemical applications. A clear application of SCL prediction is exemplified in **Chapter 4**. When applied to mass spectrometry (MS) data, SCL predictions can also be used to distinguish and cluster clinical isolates of *S. aureus* from populations with different epidemiological behaviour, causing community- or hospital-associated infections. Taken together, **Chapters 2** to **4** show how SCL predictions have already a sufficient degree of accuracy and robustness to be relevant, not only for biochemical studies, but also for the provision of valuable insights that can be translated into clinical applications that range from diagnostics to antimicrobial therapy.

Chapter 5 showcases the biological insights that can be uncovered by understanding a ML model. It was in fact for the first time shown to be possible to approximate in a quantitative manner the many factors that determine the efficiency of secretion directed by different SPs. This has for a long time been recognized as a complicated problem, and previous attempts with different approaches[13,14,16] yielded only few insights into the relevant features of SPs and how to possibly optimize them. Instead, with the interpretation analysis employed as described in **Chapter 5**, it became finally possible to explain the efficiency of each SP, to vary the SP efficiency on demand, and to come up with explanations that are in accordance with previous qualitative insights. Even though the presented model is most likely not generalizable for species other than *B. subtilis*, and for growth conditions that differ from the applied ones, it already detects the most relevant general features that influence the efficiency of a SP, including its hydrophobicity (overall and in the H-region), its charge (overall and in the three separated SP regions), the necessity of a helix-breaking residue at the edge between the H- and C-regions, and the distance of the cleavage site from its consensus sequence. Additionally, it was possible to determine other physico-chemical properties of the SP that, despite not being linked to an improved secretion could, instead, reduce SP efficiency or completely block its function. In particular, this may explain why previous studies that took into account too few factors were not successful. Thus, it seems that certain features may not have a big impact on the secretion efficiency directed by an SP but, when their value is not close to or within its optimal range, they can completely impair the SP function. Actually, taking into account all the physico-chemical features analysed, only a few SP features can be used to actually improve the protein secretion efficiency, while modification of most other features is more likely to negatively affect SP efficiency. Lastly, one important aspect that was taken into account for the first time is the possible interaction between features. Previous studies have shown that the fusion of a particular SP to different mature proteins resulted in different secretion efficiencies depending on the SP-mature protein fusion. To explain and predict such behaviour, it will be necessary to take into account also the features of the mature protein. While such an approach was so far not feasible, it will soon be completely practicable. In **Chapter 5**, the focus was nevertheless on both the interactions between different features, including both the nucleotide and amino acid contexts. Possibly because of the design that was employed, interactions

between features were found to play only a minor role, which was quite unexpected. However, it is clear that interacting features can partially compensate each other's deficiencies, or exacerbate each other's negative effects, while never overcoming the main effects of the respective features. Since only pairwise interactions were captured, another possibility is that interactions between features occur at orders higher than the second order, and it was unfortunately not possible to determine the respective values. Given that second order interactions cannot entirely explain the variability of a single physico-chemical feature, the possibility of higher order interactions should be more thoroughly investigated in the future. Furthermore, the developed model returns the mRNA secondary structure tendency at the junction between the SP and the mature protein and the one of the C-region only as the 14th and 20th most impactful features, respectively. This is possibly due to the absence of variance in the nucleotide and amino acid context of the SP. In contrast and surprisingly, the mRNA secondary structure tendency between the 5'-untranslated region (UTR) and the SP does not seem to have much impact on SP efficiency. While this is not a definite proof, the present findings suggest that also the nucleotide sequences should be taken into account when studying secretion efficiency, in contrast to previous speculations. Because the nucleotide context of the SPs was constant by design, it is not possible to determine whether it also played a role in determining protein secretion efficiency. Thus, it would not be surprising if, for instance, the pro-region had an impact on secretion not only at the amino acid level[27], but also at the nucleotide level. Additionally, there may be 'interactions' occurring between the SP and the mature protein, at least at the nucleotide level. Given the protein translocation mechanism employed by the Sec machinery, interactions between the SP and the mature protein at the amino acid level are a bit difficult to imagine, though not impossible, for co-translational translocation mediated by SRP. Altogether, the combined predictivity and interpretability of the model as presented in **Chapter 5** was demonstrated to be sufficient, both for fine tuning SPs to direct protein secretion with desired efficiency levels, and for selecting a group of best performing SPs from a list of pseudo-randomly designed SPs. This implies that a quantitative understanding of relevant physico-chemical features of SPs is emerging, and that it will soon be possible to extrapolate it towards any experimental condition.

The final experimental **Chapter 6** presents observations showing that Pro-peptides (Pros) do not necessarily improve the secretion efficiency of the protein they are fused to, as was previously proposed in some studies[30,31]. In fact, this conclusion is in agreement with most of the literature[28,29]. In particular, **Chapter 6** shows that Pros can actually reduce protein secretion levels, which is in agreement with the known chaperone and enzyme-inhibiting functions of particular Pros. However, it is conceivable that some specific features within the pro-region may influence the secretion efficiency. In fact, this would be in agreement with the view that a short synthetic Pro could actually behave more like a pro-region than as an actual chaperone, as is the case for most Pros. Such findings stress the necessity to understand the contribution of parts (e.g. the amino acid sequences of Pros) separately from their features (e.g. the charge of Pros) in order to achieve a better understanding of the underlying molecular mechanisms and their rationale. This understanding will be needed to be able to efficiently exploit Pros and pro-regions for improved protein secretion. Moreover, the results presented in **Chapter 6** show that MS data can be incredibly useful for understanding which proteases are involved in Pro cleavage, and to find out whether they are regulated through particular protease cascades. While the experimental design of these studies was not sufficiently broad to achieve a general level of understanding how the processing of Pros takes place exactly in *B. subtilis*, it was still possible to prove the concept for future, more comprehensive, investigations on the mechanisms that govern Pro processing and the function of pro-regions.

Future perspectives

Despite the revolutionary advent of integrated computational and experimental approaches in biology, many challenges still need to be overcome. For instance, with the availability of many different 'omics' technologies at reasonable costs[32] and their combined implementation with innovative microfluidic setups[33], the necessary quantitative data can be generated in sufficient amounts and with high precision to train increasingly reliable ML models. Nevertheless, the current approaches are, by and large, still inefficient with respect to resources and time. Frequently, similar data is produced multiple times due to a lack of coordination between researchers working on the same topic and, once produced, the data is not always shared in a readily usable way. Without fixed standards in the experimental design and implementation, or in data structure, the integration of data and models becomes something difficult to realize, setting severe limits to their value[32,34]. Conversely, improvements in terms of experimental design and standardisation, as well as the consideration of data re-usability right from the start of a project, will allow researchers from different scientific backgrounds to compare, test, re-analyse or even integrate the resulting data and models. The GP[4] pipeline presented in **Chapter 2** is based on these principles and, accordingly, the results can be readily used e.g. to train a ML model.

Data re-usability becomes relevant also in the realm of model interpretation, which is also referred to as explainable artificial intelligence or XAI, since previously built models could be analysed and finally explained, thereby increasing our understanding of the investigated biological system. For instance, a recent study that followed a logic opposite from the one applied in **Chapter** 5, developed a ML model to determine the best SP for a particular mature protein[35]. However, the authors did not attempt any kind of interpretation of the model. If the authors would provide free access to the model and the data used to generate it, it would now be possible for anyone to analyse and interpret this model. Additionally, although ML models are becoming more and more popular, they remain 'black-boxes' that do not provide any new insights and, thus, they do not advance our understanding of biological mechanisms. Therefore, much effort is nowadays placed in the development of either 'white-box' models or more powerful interpretation tools[7,36]. Some successful attempts have been made with respect to clinical applications[19] and drug resistance[37]. Also, easier to use and more comprehensive tools have recently been developed[38]. Taken together, it is foreseeable that model interpretation will be an important milestone on the path that biology will take with the aim to understand complex biological systems and mechanisms, thereby generating new opportunities to exploit the gathered knowledge for practical applications.

The implications of the advancement of biology towards interpretable and explainable models are astonishing. The most general and impactful advances relate to the shift from traditional trial-and-error-based wet lab approaches to approaches based on computer-assisted design (CAD). Instructive examples with industrial application potential are presented in **Chapter 5** of this thesis, and in the already mentioned study by Wu *et al.*[35]. With both ML models, it is possible to *in silico* determine the best SP that should be fused to a given protein, something that was not yet possible just 2 years ago[16]. This means that, instead of having to design, build, and screen huge libraries of SPs for each protein of interest, soon it will be possible to generate enormous libraries of SPs and screen them to choose the top 10 ones, all *in silico*. In this case, the advantage is double: while it is experimentally feasible to screen a few hundred of SP sequences, *in silico* it will be possible to screen at least thousands of SP sequences without even being restricted by the diversity present in nature. This will allow the evaluation of both bigger numbers of SPs and a higher SP diversity. Remarkably, even if the models are not 100% accurate, this still means that one needs to screen merely 10 to 20 SPs with a much higher chance of achieving a very high secretion efficiency, compared to screening hundreds of SPs with no insight into the possible outcomes.

By developing novel HT quantitative technologies, adopting new approaches, standardising them,

and by shifting towards CAD, biology will become a real engineering and industrial discipline, where important factors are known, understood, predicted, and exploitable. The impact of this paradigm shift will be incredible and ranges from boosting the bioeconomy, to resolving complex scientific questions, or optimizing specific technological approaches. In turn, this will significantly advance the fields of biomedicine, biomanufacturing, agrobiotechnology and bioremediation. In addition, fostering a spirit of collaboration and openness in science will enhance and speed up this dearly needed process of transition.

References

1. Lopatkin, A. J. & Collins, J. J. Predictive biology: modelling, understanding and harnessing microbial complexity. *Nat. Rev. Microbiol.* **18**, 507–520 (2020).
2. Angione, C. Human Systems Biology and Metabolic Modelling: A Review-From Disease Metabolism to Precision Medicine. *Biomed Res. Int.* **2019**, 8304260 (2019).
3. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-Generation Machine Learning for Biological Networks. *Cell* **173**, 1581–1592 (2018).
4. Pérez-Velázquez, J., Gölgeli, M. & García-Contreras, R. Mathematical Modelling of Bacterial Quorum Sensing: A Review. *Bull. Math. Biol.* **78**, 1585–639 (2016).
5. Succurro, A. & Ebenhöh, O. Review and perspective on mathematical modeling of microbial ecosystems. *Biochemical Society Transactions* vol. 46 403–412 (2018).
6. Birkegård, A. C., Halasa, T., Toft, N., Folkesson, A. & Græsbøll, K. Send more data: A systematic review of mathematical models of antimicrobial resistance 01 Mathematical Sciences 0102 Applied Mathematics. *Antimicrobial Resistance and Infection Control* vol. 7 1–12 (2018).
7. Yu, M. K. *et al.* Visible Machine Learning for Biomedicine. *Cell* **173**, 1562–1565 (2018).
8. Nielsen, H. Protein sorting prediction. in *Methods in Molecular Biology* vol. 1615 23–57 (Humana Press Inc., 2017).
9. Nielsen, H. Predicting subcellular localization of proteins by bioinformatic algorithms. in *Current Topics in Microbiology and Immunology* vol. 404 129–158 (Springer Verlag, 2017).
10. Sato, K. *et al.* A protein secretion system linked to bacteroidete gliding motility and pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 276–281 (2010).
11. Götz, F., Yu, W., Dube, L., Prax, M. & Ebner, P. Excretion of cytosolic proteins (ECP) in bacteria. *Int. J. Med. Microbiol.* **305**, 230–7 (2015).
12. Ebner, P., Rinker, J. & Götz, F. Excretion of cytoplasmic proteins in Staphylococcus is most likely not due to cell lysis. *Curr. Genet.* **62**, 19–23 (2016).
13. Brockmeier, U. *et al.* Systematic Screening of All Signal Peptides from Bacillus subtilis: A Powerful Strategy in Optimizing Heterologous Protein Secretion in Gram-positive Bacteria. *J. Mol. Biol.* **362**, 393–402 (2006).
14. Degering, C. *et al.* Optimization of protease secretion in bacillus subtilis and bacillus licheniformis by screening of homologousand heterologous signal peptides. *Appl. Environ. Microbiol.* **76**, 6370–6376 (2010).
15. Fu, G., Liu, J., Li, J., Zhu, B. & Zhang, D. Systematic Screening of Optimal Signal Peptides for Secretory Production of Heterologous Proteins in Bacillus subtilis. *J. Agric. Food Chem.* **66**, 13141–13151 (2018).
16. Peng, C. *et al.* Factors influencing recombinant protein secretion efficiency in gram-positive bacteria: Signal peptide and beyond. *Front. Bioeng. Biotechnol.* **7**, (2019).
17. Samek, W. Learning with explainable trees. *Nat. Mach. Intell.* **2**, 16–17 (2020).
18. Lundberg, S. M., Allen, P. G. & Lee, S.-I. I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems* (2017).
19. Lundberg, S. M. *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
20. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
21. Slack, D., Hilgard, S., Jia, E., Singh, S. & Lakkaraju, H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. in *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 180–186 (ACM, 2020). doi:10.1145/3375627.3375830.

22. Challen, R. *et al.* Artificial intelligence, bias and clinical safety. *BMJ Quality and Safety* vol. 28 231–237 (2019).
23. Jiang, H. & Nachum, O. *Identifying and Correcting Label Bias in Machine Learning*. *ArXiv* vol. cs.LG (2019).
24. Meyer, H. & Pebesma, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *ArXiv* **stat.ML**, (2020).
25. Otto, A. *et al.* Systems-wide temporal proteomic profiling in glucose-starved Bacillus subtilis. *Nat. Commun.* **1**, 137 (2010).
26. Petzold, C. J., Chan, L. J. G., Nhan, M. & Adams, P. D. Analytics for metabolic engineering. *Frontiers in Bioengineering and Biotechnology* vol. 3 135 (2015).
27. Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A. & Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *European Journal of Cell Biology* vol. 97 422–441 (2018).
28. Demidyuk, I. V., Shubin, A. V., Gasanov, E. V. & Kostrov, S. V. Propeptides as modulators of functional activity of proteases. *Biomolecular Concepts* vol. 1 305–322 (2010).
29. Takagi, H. & Takahashi, M. A new approach for alteration of protease functions: pro-sequence engineering. *Appl. Microbiol. Biotechnol.* **63**, 1–9 (2003).
30. Kouwen, T. R. H. M. *et al.* Contributions of the Pre- And Pro-Regions of a Staphylococcus hyicus Lipase to Secretion of a Heterologous Protein by Bacillus subtilis. *Appl. Environ. Microbiol.* **76**, 659–669 (2010).
31. Sturmfels, A., Götz, F. & Peschel, A. Secretion of human growth hormone by the food-grade bacterium Staphylococcus carnosus requires a propeptide irrespective of the signal peptide used. *Arch. Microbiol.* **175**, 295–300 (2001).
32. Manzoni, C. *et al.* Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief. Bioinform.* **19**, 286–302 (2018).
33. Bai, Y. *et al.* Applications of Microfluidics in Quantitative Biology. *Biotechnol. J.* **13**, 1700170 (2018).
34. Li, Y., Wu, F. X. & Ngom, A. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics* vol. 19 325–340 (2018).
35. Wu, Z. *et al.* Signal Peptides Generated by Attention-Based Neural Networks. *ACS Synth. Biol.* acssynbio.0c00219 (2020) doi:10.1021/acssynbio.0c00219.
36. Azodi, C. B., Tang, J. & Shiu, S. H. Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends in Genetics* vol. 36 442–455 (2020).
37. Yang, J. H. *et al.* A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. *Cell* **177**, 1649-1661.e9 (2019).
38. Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability. *ArXiv* **cs.LG**, (2019).

APPENDIX

A

LIST OF PUBLICATIONS

List of publications

1. Grasso, S., van Rij, T. & van Dijl, J. M. GP4: an integrated Gram-Positive Protein Prediction Pipeline for subcellular localization mimicking bacterial sorting. *In press for Briefings in bioinformatics,* (2020).
2. Gabarrini, G., Grasso, S., van Winkelhoff, A. J. & van Dijl, J. M. Gingimaps: Protein Localization in the Oral Pathogen *Porphyromonas gingivalis*. *Microbiol. Mol. Biol. Rev.* **84**, (2020).
3. Nepal, S. *et al.* An ancient family of mobile genomic islands introducing cephalosporinase and carbapenemase genes in *Enterobacteriaceae*. *Virulence* **9**, 1377–1389 (2018).
4. Mekonnen, S. A. *et al.* Signatures of cytoplasmic proteins in the exoproteome distinguish community- and hospital-associated methicillin-resistant *Staphylococcus aureus* USA300 lineages. *Virulence* **8**, 891–907 (2017).
5. Beier, S. *et al.* Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Sci. Data* **4**, 170044 (2017).
6. Mascher, M. *et al.* A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**, 427–433 (2017).
7. Grasso, S. & Tell, G. Base excision repair in Archaea: back to the future in DNA repair. *DNA Repair (Amst).* **21**, 148–57 (2014).
8. Colitti, M. & Grasso, S. Nutraceuticals and regulation of adipocyte life: premises or promises. *Biofactors* **40**, 398–418 (2014).