PATTERN SEARCH IN PATHOGENIC BACTERIAL PROTEINS FOR LOCALIZATION AND SECRETORY SYSTEMS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY
ORHAN ÖZCAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
BIOTECHNOLOGY

AUGUST 2015

Approval of the thesis:

PATTERN SEARCH IN PATHOGENIC BACTERIAL PROTEINS FOR LOCALIZATION AND SECRETORY SYSTEMS

submitted by **ORHAN ÖZCAN** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Biotechnology, Middle East Technical University** by,

Prof. Dr.Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**
Prof. Dr. Filiz Bengü Dilek _____
Head of Department, **Biotechnology**
Prof. Dr. Gülay Özcengiz _____
Supervisor, **Biology Dept., METU**
Assoc. Prof. Dr. Tolga Can _____
Co-Supervisor, **Computer Engineering Dept., METU**

Examining Committee Members:
Prof. Dr. Haluk Hamamcı _____
Food Engineering Dept., METU
Prof. Dr. Gülay Özcengiz _____
Biology Dept., METU
Assoc. Prof. Dr. Yeşim Aydın Son _____
Informatics Dept., METU
Prof. Dr. Osman Uğur Sezerman _____
BioStatistics Dept., ACU
Assoc. Prof. Dr. Servet Özcan _____
Genome and Stem Cell Center., ERU

Date: 27.08.2015

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Orhan ÖZCAN

Signature:

ABSTRACT

PATTERN SEARCH IN PATHOGENIC BACTERIAL PROTEINS FOR LOCALIZATION AND SECRETORY SYSTEMS

Özcan, Orhan

Ph.D., Department of Biotechnology

Supervisor: Prof. Dr. Gülay Özcengiz

Co-Supervisor: Assoc. Prof. Dr. Tolga Can

August 2015, 184 pages

Computational prediction of bacterial protein localization (BPL) is a very useful tool which provides clues about protein function. For pathogenic proteins in particular, detection of their subcellular location and their secretory pathways have great implications for vaccine and drug design. Cell surface and/or secreted proteins of microbes can also be used as biomarkers for sensor applications. At present, there are numerous BPL prediction algorithms and programs available, however, most of them give false positive results in order to maximize the number of positive predictions. Moreover, state of the art algorithms, specifically PSORT, successfully identify protein localization for every organism from any given sequence information but they usually fail in pathogenic sequences. Because the most of the pathogenic proteins are surface-localized, there is an imminent need for pathogen-specific secretion motif search algorithms as well. These motifs would also provide information on bacterial protein localization.

In the present work, we built databases of pathogenic sequences and searched for selected 5 to 18 amino acid long motifs as a new approach, namely Pathogenic Sequence Motif Search (PSMS). The algorithm is based on a total of 52 distinct secretion-associated patterns covering 6 different secretory pathways for the prediction of surface and secreted proteins. The datasets for each of the following groups of proteins were next established for our validation studies which involved the tests for the success rate of these 52 patterns: Secreted, immunoreactive and patented vaccine, cytoplasmic and orphan-secreted with 3241, 1740, 2582 and 2533 members, respectively. A total of 3241 proteins in secreted proteins dataset represented TISSS, T2SS, T3SS, T4SS, T5SS and T6SS systems of secretion with 954, 668, 381, 770, 221 and 274 protein sequences, respectively. Cytoplasmic protein dataset, on the other hand, was used to exclude certain candidate patterns. 43 out of 52 patterns were truly secretion-related, pointing directly to a specific secretion system. Rest 9 patterns were found in secreted proteins though not related to a specific secretion system. Additionaly, LC-MS data formerly obtained in our laboratories from *Bordetella pertussis* surface proteome and secretome analyses were also included in the secreted protein sequence dataset. The selected patterns were demonstrated for instance in 503 out of a total of 1740 proteins in the immunoreactive protein dataset.

With the help of our patterns, 75 proteins which were formerly predicted to have an intracellular localization and mistakenly ruled out as potential drug targets/vaccine candidates were successfully predicted as surface-associated/secreted ones. Besides the development of PSMS program predicting pathogenic sequences with high accuracy, the separate databases constructed in this work with respect to immunoreactivity and distinct secretory pathways are expected to constitute valuable bioinformatics resources for researchers of the field.

Keywords: Bacterial subcellular localization prediction algorithms, Pathogenic Sequence Motif Search (PSMS), Pathogenic protein motifs,, Surfacome, Protein secretion systems, Protein sequence databases

ÖZ

PATOJENİK BAKTERİYEL PROTEİNLERDE SALGI SİSTEMLERİ İÇİN ÖRÜNTÜLER ARANMASI

Özcan, Orhan
Doktora, Biyoteknoloji
Tez Yöneticisi: Prof. Dr. Gülay Özcengiz
Ortak Tez Yöneticisi: Doçent. Dr. Tolga Can
Ağustos 2015, 184 sayfa

Bilgisayar temelli bakteriyel protein lokalizasyon (BPL) öngörüsü, proteinlerin fonksiyonları hakkında bilgiler veren çok kullanışlı bir araçtır. Özellikle patojenik proteinlerin hücresel alt lokasyonlarının ve salgılandıkları yolakların anlaşılması, potansiyel ilaç ve aşı hedeflerinin ortaya çıkarılması ve hatta mikroorganizmalar için sensor uygulamalarında kullanılabilecek biyomarkörlerin geliştirilmesi için çok önemlidir. Günümüzde birçok BPL öngörü algoritması ve programı mevcut olmasına rağmen bunların birçoğu olumlu öngörü sayısını azami seviyede tutmak amacıyla hazırlandığından hatalı pozitif sonuçlar vermektedir. Bunun yanısıra, mevcut güncel programlar, örneğin PSORT, prensip olarak her organizma türü için verilen sekansları başarıyla analiz edip hücresel localizasyon öngörüsü yapabilmekte, ancak patojenik sekanslar için genellikle başarısız olmaktadır. Patojenik proteinlerin çoğu bakterinin yüzeyinde lokalize olduğundan, patojene spesifik sekresyon motif tarama algoritmalarına da büyük gereksinim vardır ve bunların temelini oluşturan motifler aynı zamanda bakterilerde protein lokalizasyonu hakkında da bilgi verecektir.

Şimdiki çalışmada, patojenik bakteriyel protein sekansları içeren veri kümeleri oluşturulmuş ve seçillmiş 5-18 amino asit uzunluğunda motifler taranarak "Pathogenic Sequence Motif Search (PSMS).isimli yeni bir algoritma geliştirmiştir. Bu algoritma, yüzey proteomu ve sekretom komponentleri öngörüsü için 6 farklı sekresyon yolağına karşılık gelen toplam 52 salgılama ile ilişkili protein kalıbının seçilip kullanılmasını temel almaktadır. Bu kalıpların (i) salgılanan, (ii) immünoreaktif ve patentli aşı komponentleri, (iii) sitoplazmik ve (iv) orfan-salgılanan protein grupları için oluşturulan veri kümelerinde taranarak başarı oranlarının test edilmesini içeren doğrulama çalışmaları yapılmıştır. Bu veri kümeleri, yukarıda verilen sıraya göre 3241, 1740, 2582 ve 2533 üyeye sahiptir. Salgılanan proteinleri içeren veri kümesinde mevcut 3241 protein sekansı, sırasıyla TISSS, T2SS, T3SS, T4SS, T5SS ve T6SS salgılama sistemlerini temsil eden 954, 668, 381, 770, 221 ve 274 protein sekansının toplamıdır. Daha önceki çalışmalarımızda *Bordetella pertussis* yüzey proteomu and sekretomundan elde edilen LC-MS bulgularımız da salgılanan protein sekansı veri kümesinde kullanılmıştır. Seçilmiş kalıpların varlığı, örneğin immünoreaktif proteinleri ve aşı komponentlerini içeren veri kümesinde mevcut toplam 1740 proteinden 503'ünde doğrulanmıştır. Sitoplazmik veri kümesi ise uygun olmayan kalıp adaylarını dışlayabilmek amacıyla kullanılmıştır. Validasyon çalışmaları, mevcut 52 kalıptan 43'ünün salgılama sistemleri ile doğrudan ilişkili, geri kalan 9 kalıbın ise salgılanan, ancak salgılamanın spesifik bir sistemi ile ilişkilendirilemeyen proteinleri belirlediğini göstermiştir.

Bu kalıpların yardımıyla, daha önce PSORT kullanılarak hücre içi lokalizasyona sahip olduğu gösterilmiş ve bu nedenle ilaç hedefi/aşı adayı olamamış 75 ayrı proteinin aslında hücre yüzeyinde/salgılanan proteinler olduğu başarıyla öngörülmüştür. Patojenik proteinleri yüksek doğrulukla öngören PSMS programı ve bu tip proteinler için immünoreaktivite ve sekresyon sistemleri temelinde oluşturulan verikümeleri ilgili alandaki araştırmacılar için değerli biyoinformatik kaynaklar oluşturacaktır.

Anahtar Kelimeler: Bakteriyel hücresel lokalizasyon öngörüsü, Patojenik sekans motif arama (PSMS), Patojenik protein motifleri, Protein sekresyon sistemleri, Protein sekans databazları

To My Wife,

ACKNOWLEDGMENTS

It would not have been possible to write this doctoral thesis without the help and support of the upholder people around me, to only some of whom it is possible to give particular mention here.

I must offer my deepest gratitude to my thesis advisors, Prof. Dr. Gülay Özcengiz, Prof. Dr. Uğur Sezerman and Assoc. Prof. Tolga Can. They offered their unreserved help and guidance and led me to finish my thesis step by step. Their words inspired me and elevated me to a higher level of thinking, viewing this world from a new perspective. Without their kind and patient instruction, it is impossible for me to finish this thesis. Moreover, I want to give special gratitude to Prof. Dr. Uğur Sezerman. He gave me valuable suggestions and enlighten the pathway when the thesis struggles with various unresolved challenges. I also want to thank 118 Lab members at TUBITAK GMBE; Assoc. Prof. Yavuz Öztürk, İbrahim Sertdemir, Sevde Şencan and Assoc. Prof. M. Naci Yazıcıoğlu who helped correcting my thesis carefully and offered me inspiring suggestions in the oral defense.

I would like to thank my wife Aysun for her personal support and great patience at all times. My parents have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice.

Amongst my fellow postgraduate students in the Department of Biological Sciences, the effort made by Lab-214 and Lab 207 members in promoting a stimulating and welcoming academic and social environment will stand as an

example to those that succeed them. Especially with Lab-214 members we inhabited whole life in one laboratory. I thankfully acknowledge my 207 and 214 labmates: Dr. Aslıhan Kurt, Ayça Çırçır, Çiğdem Yılmaz, Elif Tekin, Mustafa Demir, Mustafa Çiçek, Alper Mutlu, Eser Ünsaldı, İsmail Cem Yılmaz, Dr. Sezer Okay, Dr. Volkan Yıldırım and Dr. B. E. Tefon. for their friendship and cooperation.

I would like to thank Prof. Dr. Haluk Hamamcı, head of jury, and jury members; Assoc. Prof. Dr. Servet Özcan, Assoc. Prof. Dr. Yeşim Aydın Son and Assoc. Prof. Dr. Mehmet Somel for evaluation of thesis project and found me qualified for PhD degree.

I would like to thank Gökalp Çelik for helping multi-threading facility. I would like to thank Seren Sert for several code improvements in handling helper classes. I would also thank MSDN and Microsoft for supporting debugging errors. I am very pleased to use Visual studio 2013 student edition with its powerful, epic, legendary debugging.

Finally, I would like to acknowledge the Middle East Technical University and its staff, particularly in hiring of a Research Assistantship that provided the necessary financial support for this research.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVATIONS

AntigenDb	Antigen Database
CD-HIT	Cluster Database at High Identity with Tolerance
HMM	Hidden Markov Model
k-mer	all the linear possible subsequences (of length k)
NCBI PDB	National Center for Biotechnology Information Protein Data Bank
PROSITE	Database of protein domains, families and functional sites
ProRule	A collection of PROSITE rules based on profiles and patterns
SVM	Support Vector Machine
SPC	Short Pattern Candidate
TISSS	Type 1 Secretion System
T2SS	Type 2 Secretion System
T3SS	Type 3 Secretion System
T4SS	Type 4 Secretion System
T5SS	Type 5 Secretion System
T6SS	Type 6 Secretion System
T7SS	Type 7 Secretion System
T8SS	Type 8 Secretion System
Uniprot	The Universal Protein Resource
Uniprot UniRef	The UniProt Reference Clusters
USPTO	United States Patent Trademark Office

CHAPTER 1

INTRODUCTION

1.1. Thesis Statement

In the present study, we took the advantage of the availability of pathogen surfacome and secretome data and attempted to construct the first inductive systematic localization-associated pattern finding software with the aim of providing further insight into localization prediction. Our approach involved the computational analysis of non-cytoplasmic proteomic data. For the identification of novel secretion associated patterns, we developed a motif search algorithm. We also established several datasets including one from our own *Bordetella pertussis* LC-MS data.

1.2. Motivation

The need for development of a new program for localization prediction emerged as soon as we performed an immunopreoteomic analysis of *B. pertussis*. In order to determine proper vaccine candidates, immunological proteins are filtered out by several bioinformatics tools. In our previous immunoproteomic analysis, PSORTb (Yu et al., 2010), PrediSi (Hiller *et al*., 2004) and SignalP (Nordahl *et al*., 2011) were used as localization predictors. When searching for vaccine candidates, several higher abundant surfacome

proteins may be predicted as cytoplasmic proteins, and therefore, be eliminated from further immunological tests. For example *groL* 60 kDa has been one of the eliminated proteins for further analysis due to its predicted cytoplasmic localization. Since the identification of this chaperonin in 1988, the protein has been identified in various pathogen secretome and surfacome analyses. . The first scientific work about possible localization of this chaperonin on pathogen surface was published in 2006 (Frisk *et al*., 2006). Because of the common localization knowledge, the predictor's cytoplasmic localization prediction, and due to the below report, the protein was not favored as a vaccine candidate before 2010. Yamaguchi *et al*., (1997) reported that there is a common epitope in *H. pylori* Hsp60 and human gastric epithelial cells. The elimination of *groL* 60kDa from being vaccine candidate for about 13 years has been the leading motivation in this study. Another example of a misleading localization prediction of the current programs is the protein AvrBs3. It is predicted as a cytoplasmic protein due to its DNA binding domain. In reality, the protein is an effector protein which is secreted by a Type 3 secretion system (Paul Dean, 2011). Likewise, there might be other proteins, which cannot be identified as vaccine candidates, due to their misleading localization prediction. An incorrect prediction is possible even when a protein has been studied over decades. Although the prediction programs are highly sophisticated and produce useful prediction, most of them generate false-positive results in order to increase sensitivity.

Most localization prediction programs increase their success rate by using the function information of domains on the protein. Although it is a successful predictor for most of the proteins, it may cause misleading localization results for pathogenicity-related proteins since functions of the domains on these proteins are related to target organisms rather than the producer. For example; a pathogenic protein with an effect on the cytoplasm of target organism is actually a secreted protein of the producer, not a cytoplasmic protein; but, prediction programs determine its location as cytoplasm due to the presence

of the domain and its action in target cytoplasm. In this study, the Pathogenic Sequence Motif Search (PSMS) was developed to tackle this problem as a de-novo protein motif finder and analysis tool.

1.3. Secretion Systems in Bacteria

None of the proteins, enzymes and toxins are waste materials for bacterial secretion systems and they all possess crucial functions for both survivability and pathogenicity (Delepelaire, 2004). There are several pathways for bacterial secretion such as twin arginine translocation system, Sec protein secretion, type I, II, III, IV and V secretion systems, and also non-classical secretion systems of Gram positive and Gram negative pathogens (Wooldridge, 2009).

1.3.1. Type 1 Secretion System (TISS)

As the simplest secretion system, it is present in both animal and human pathogens (Jenewein, et al., 2009). Three groups of proteins are involved in Type 1 secretion systems; ATP-binding cassette (ABC), Membrane Fusion Protein (MFP) and specific outer membrane protein (OMP) (Figure 1.1 and Figure1.2). As it is Sec-independent, it bypasses periplasm (Delepelaire, 2004). Many of the Type 1 secreted proteins carry glycine-rich repeats (**GGXGXDXXX**) that specifically bind calcium ions (Delepelaire, 2004). Proteins of various size can be transported with the presence of ABC protein recognition signal. ABC protein recognition signal mostly resides on C-terminal region of protein rather than N- terminal region. However, Masi & Wandersman (2010) reported that multiple signals might be involved in T1SS.

1.3.2. Type 2 Secretion System (T2SS)

Transportation to periplasm with Sec or Tat system is required for Type 2 (T2SS) secreted proteins. N-terminal sequence tag of protein is needed and as soon as the protein passes through cytoplasmic membrane, the tag is cleaved. Proteases, cellulases, pectinases, phospholipases, lipases, and some other

proteins like toxins whose concentrations are dependent on quorum-sensing are T2SS proteins and their presence is crucial for virulence (Sandvist M., 2001). The signal peptides that are needed for Sec or Tat system possesses N-, H-, and C region. N-region is located on the N-terminus and it is positively charged. H-region is hydrophobic region and resides in membrane. C-region, bearing the characteristic cleavage site, is generally small and uncharged (Von Heijne G., 1990). It is shown that many signal peptides of T2SS proteins are conformational signals (The 3D structure of protein itself provides a signals called conformational signals) and folding is prerequisite for transportation to extracellular space (Chapon et. al., 2001). Voulhoux et. al showed that, the deletion of the domains of *Pseudomonas aeruginosa* Exotoxin A leads to identify specific signals that T2SS might be involved with conformational motif (Voulhoux et. al., 2000). The proteins folded by chaperons in periplasm exhibits conformational signals for transportation to extracellular space. (Voulhoux et. al., 2000) (Figure 1.2 b).

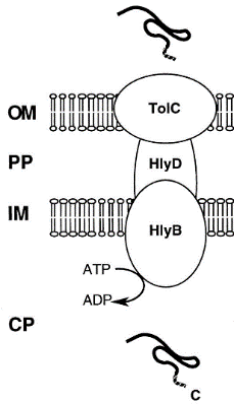


Figure 1.1. Schematic exemplification of TISS by alpha-hemolysin secretion in *E. coli* (Hueck, 1998). CP, cytoplasm; IM, inner membrane; PP, periplasm and OM, outer membrane.

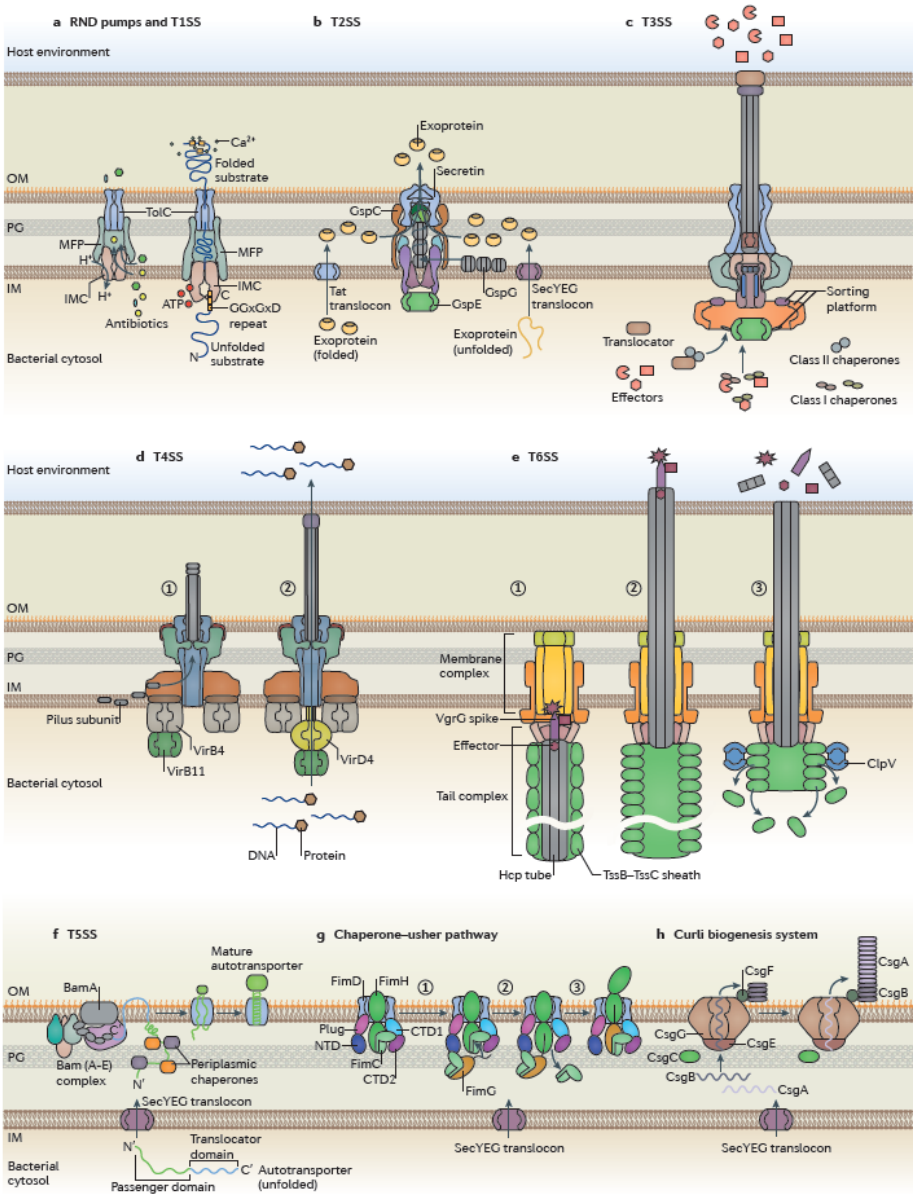


Figure 1.2. Secretion models of gram-negative bacteria (Tiago et al., 2015). a- TISS. b-T2SS. c-T3SS. d-T4SS. e-T6SS. f- T5SS. g- T7SS. h-T8SS.

1.3.3. Type 3 Secretion System (T3SS)

With respect to existence in pathogens of distant evolutionary lines, T3SS is conserved in four major unrelated plants and several animal pathogenic genera. With the genetic analysis of the virulent and avirulent strains of the organisms revealed that 20 genes involved in T3SS (Hueck CJ., 1998). By means of T3SS, not only virulence factors, similar to the host signal transduction elements, but also symbiotic elements are transported into the host cells. T3SS is involved in the flagella biosynthesis, and just similar to flagella biosynthesis, molecular needles (Figure 1.3.) are constructed and signal transduction elements are transferred to the host cells through the needles. With its complex structure and function, it can be counted as a bacterial organelle that has specifically evolved to deliver bacterial effector proteins (Gala´n and Wolf-Watz, 2006). Multiple signals can be seen for substrate specificity and organisms may express more than one T3SS to ensure correct targeting (Gala´n and Wolf-Watz, 2006).

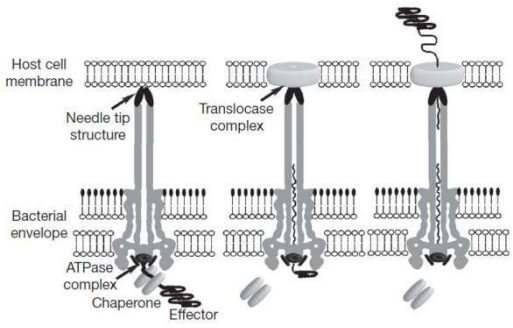


Figure 1.3. Illustrations of Gala´n and Wolf-Watz, (2006) showing molecular syringe of T3SS.

T3SS signals can be seen in protein sequence, mRNA sequence, both of the protein and mRNA sequences or none of protein and mRNA sequences. The coding mRNA signal found with the help of a frame-shift mutation insertion on *Yersinia enterocolitica* (Andersonand Schneewind, 1997)**.** Identified non-cleaved amino acid signals of T3SS are mostly located in first 30 amino acids of proteins. Prior to insertion inside the molecular needle, effector protein should be transformed into a secretion competent structure with the help of the chaperones. Protein should be turned into non –globular polypeptide sequence. T3SS-associated chaperones control the secretion specificity (Lee & Galán, 2004)

1.3.4. Type 4 Secretion System

Unlike the other secretion systems discussed previously, Type 4 secretion (T4SS) is not only capable of transporting proteins but also capable of transporting DNA. (Lessi et al., 1992). In terms of the function and the structural formation, T4SS is divided in three subdivisions: (i) Translocation of single stranded DNA for the conjugation process while contacting with other cell, (ii) delivering substrates into eukaryotic cells, (iii) taking up DNA and proteins from extracellular environment or releasing DNA or proteins into the extracellular space.4S (Trokter et. al., 2014).

Instead of single signals, bipartite signals can be also seen in T4SS (Schulein et. al., 2004). Global alignments of several bacterial classes of T4SS effector proteins for revealing motif candidates and short amino acid sequence usage probabilities were demonstrated by Wang et al., 2014.

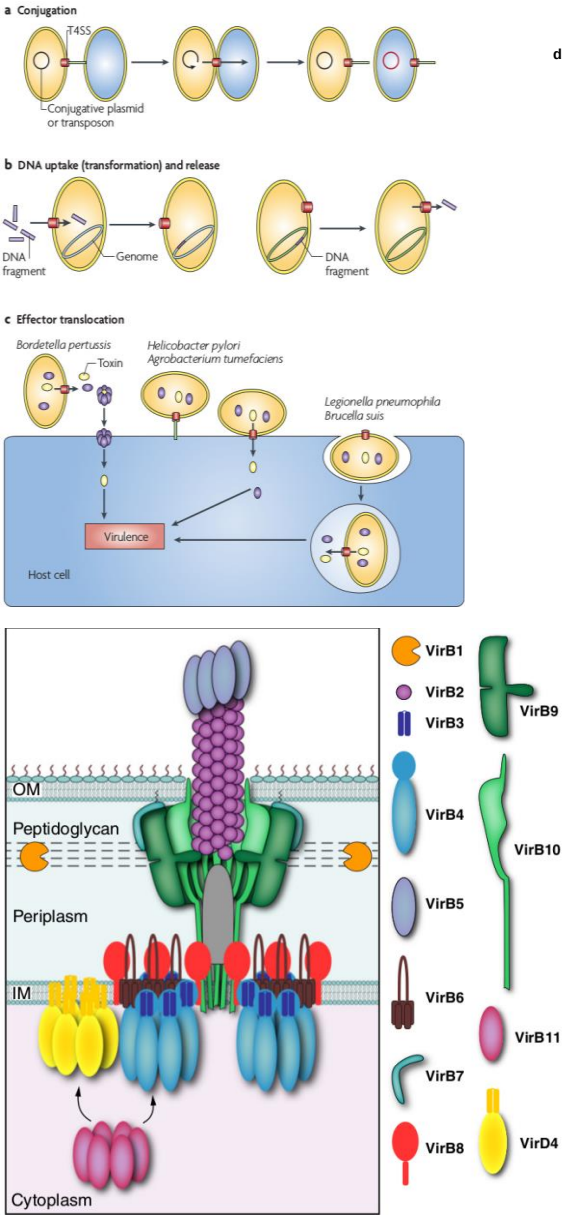


Figure 1.4. Schematic representation of all T4SS subdivisions and T4SS elements. **a**| contact required DNA translocation (conjugation). **b**| Transformation where contact is not required. **c**| Virulence factor translocation (Frozes *et. al*., 2009). **d**|Schematic representation of T4SS elements (Trokter *et. al*., 2014).

1.3.5. Type 5 Secretion System (T5SS)

Like T2SS, autotransporter or type 5 secretion (T5SS) requires Sec system for passing through inner membrane (Thanassi *et. al.*, 2005). Proteins that are secreted via this way have similarities in primary sequences (Henderson et. al., 2004). T5SS is divided in three subpathways. Among these, T5SSa serves for protease secretion in pathogens (Pohlner *et. al.*, 1987). T5SS secreted effector proteins are unfolded in both cytoplasm and in periplasm of pathogens so that they can escape from self-toxicity of effector molecules.

IgA protease of *Neisseria gonorrhoeae* is secreted by T5SSa system. Protease is transferred from cytoplasm to periplasm with Sec system in an unfolded state (Pohlner *et al.*, 1987). This protein is secreted outside of the cell from periplasm in the absence of ATP (Thanassi *et. al.*, 2005).

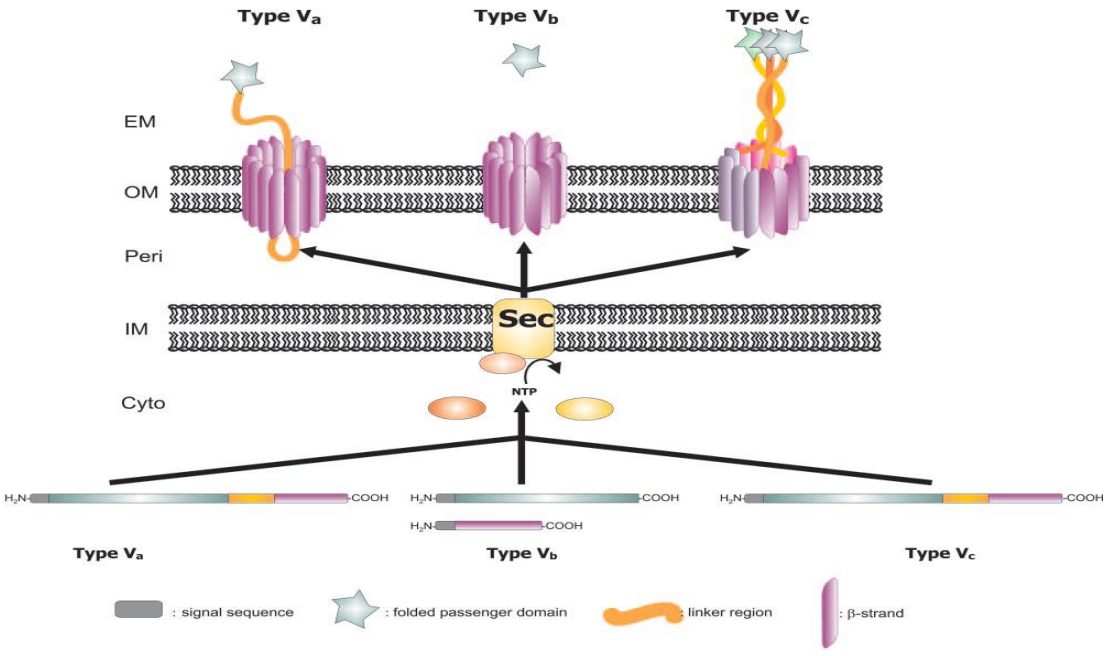


Figure 1.5. Schematic overview of T5SS of Henderson *et. al.*, 2004.

1.3.6 Type 6 Secretion System (T6SS)

Almost ¼ of all proteobacterial genomes possesses this type of secretion system (Bigle *et al*., 2008). Although the most of the studies on Type 6 secretion system is related with its pathogenicity roles, this secretion system is also used as a defense system for protozoal predators (Coulthurst *et al*., 2013) The study on T6SS system of *Vibrio cholera* showed that the pathogen uses this secretion system as a sword to burst other intestinal bacteria aligning near intestines. The antimicrobial activity of T6SS (threonine phosphorylation pathway: TPP) in *P. aeruginosa* attracks scientists' attention (Coulthurts et al., 2013). The precise signal sensed by the TPP remains to be reported.

1.3.7. Type 7 Secretion System (T8SS)

This type of secretion system is specific to *Mycobacterium*. Evolutionary relative of the Type 4 secretion is ESX or Type 7 secretion (Bitter et al. 2009) The loss of RD1 region in *Mycobacterium* may contribute to the development of attenuated vaccines (Pym *et al*., 2002) since T8SS is controlled by the RD1 region of the *Mycobacterium* and currently 4 secreted proteins were identified (Pym *et al*., 2002 and Xu *et al*., 2007). According the work done by Lukra *et al*. (2008), 10 ORFs reside in RD1 locus. The authors stated that Rv3870 and Rv3876 (2 ORF in RD1) are ATPases so that T8SS might work in an ATP-dependent manner. How many proteins or gene products secreted was not clear in literature so that we search UniProt database using the ORFs as defined by Lukra *et al*. (2008) as a template and totally 1583 CDD and 3396 source hits captured with CDHIT 0.5 similarity.

1.3.8 Type 8 Secretion System

The "curli" pathway or the extracellular nucleation-precipitation is the Type VIII secretion system (T8SS) in Gram-negative bacteria. Type 8 is responsible for the secretion and assembly of the prototypical curli. The curli is actually aggregative fimbriae that their secretion held by T8SS. Curli was

first identified in *Salmonella* spp (Gibson et al., 2007) Gibson showed curli is under the control of agf operon.

1.4. Proteomics for Biopharmaceuticals Industry

The term proteome refers to all proteins expressed by an organism at a specific time and space. The fast improvements in genomics stimulated proteomics as the identification of proteins by mass spectrometry depends on genome-derived theoretical protein cleavages by proteases. (Wilkins et al., 1996). Soon after the protein isolation from the sample organism/tissue, the proteins are separated according to the certain physiochemical properties which is called fractionation. Separated proteins are turned into peptides by tryptic digestion. Tryptic digested peptides are measured with the help of a mass spectrometer and the proteins are identified by searching the databases. A short explanation of these steps is presented in Figure 1.6.

Although biopharmaceuticals industry (antibodies and other protein drugs such as hormones, growth factors, cytokines, enzymes, etc.) is growing fast, the search for novel, more effective, and safer drugs is still the mainstay of current pharmaceutical research. Understanding target at molecular level is the basis of modern drug discovery. (Congreve et al., 2005, Rondeau & Schreuder, 2008). Structural biochemistry and the proteomics developments has significantly accelerated drug discovery process.

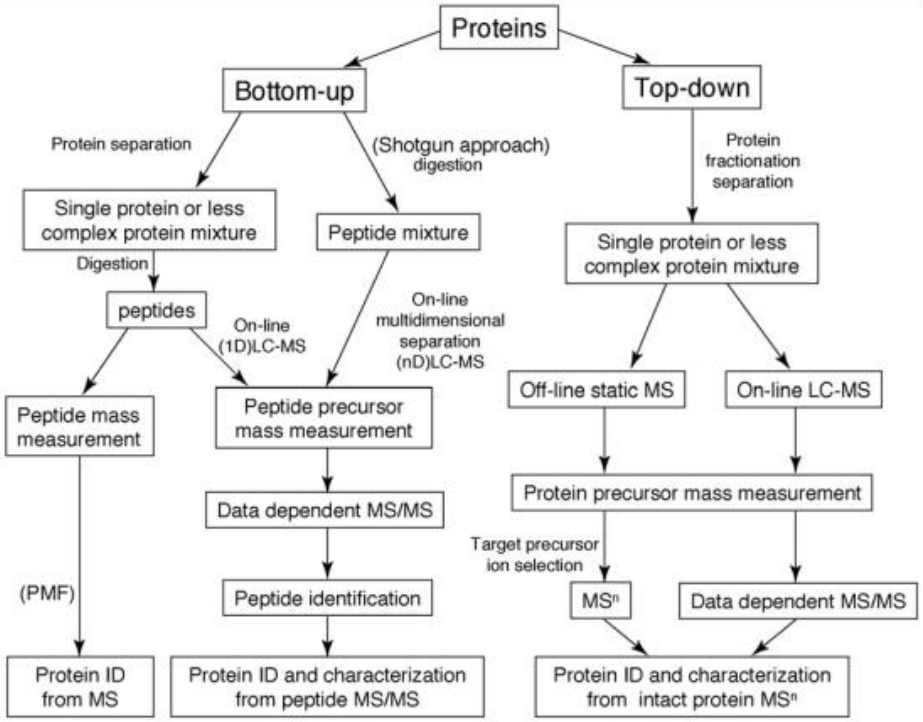


Figure 1.6. Schematic explanation of proteomic strategies of proteomics (Han, 2008).

Bacterial genomes are considered to be coding a small set of proteins relative to higher organisms. However, wet-lab experiments for bacterial vaccine development consist of many steps which are laborious and time consuming. Vaccine technology greatly benefits from immunoproteomics, which studies the immune-reactive protein sets of proteome. Immunocapturing, 2D Western gel electrophoresis, high-throughput peptide and protein synthesis coupled with ELISPOT analysis provide valuable information for immunoproteomics. Even with the advanced immunoproteomics tools, great amount of work is required for determining vaccine candidates. Such 18 examples is seen in the reviews of Velayudhan and Porcelli (2013), who summarized 18 different researches consisting nearly 500 cloned genes to find vaccine candidates

against *Mycobacterium tuberculosis* (*Mtb*). Figure 1.7 summarizes the *Mtb* peptidome and proteome-wide screening of *Mtb* where billions of peptides and thousands of proteins were analyzed, and there is no effective vaccine for all forms of tuberculosis (TB). This clearly shows the need for more effective filtering of the immunoproteomic data with better bioinformatics tools.

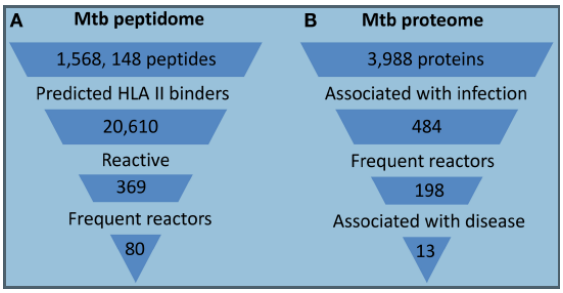


Figure 1.7. The summary of the proteome-wide screens for *Mtb* antigens (Velayudhan and Porcelli (2013). a- Summary of screen for target peptides of CD4 CT cells b- Summary of target proteins of humoral responses.

1.5. Literature of Localization Prediction

In order to provide necessary information and clues about the function of a protein, it is important to predict its subcellular localization (SCL) from its amino acid sequence (Petsalaki *et al*., 2006). Although, there are wet-lab methods, such as green fluorescent protein (GFP) tagging (Sawin & Nurse, 1996) and gene trap screening (Sutherland *et al*., 2001), which provide localization information of proteins, there is still a need for fast, automated, cheap, and accurate approaches for predicting protein localization for different kinds of cells especially when there is massive proteome data in hand. There are numerous bacterial localization prediction (BLP) methods and most of them produce false-positive results in order to have better

sensitivity. PSORT is the first program to predict localization (1991). More than 40 different predictors have been released in the last 25 years (http://www.psort.org/). The widely used prediction programs are listed in Table 1.1.

Table 1.1. Widely used protein localization prediction programs.

Name	Reference	Name	Reference
Cell-PLoc	Chou KC and Shen HB, 2008	PSORTb	Gardy et. al., 2003
BaCelLo	Pierleoni et. al., 2006	MetaLocGramN	Magnus et. al., 2012
CELLO	Yu et. al., 2004	PredictNLS	Nair et. al., 2003
ClubSub-P	Paramasivam and Linke, 2011	SecretomeP	Bendtsen et. al., 2004
Euk-mPLoc 2.0	Chou and Chen, 2010	SherLoc	Shatkay et. al., 2007
CoBaltDB	Goudenege et. al., 2010	TargetP	Emanuelsson et. al., 2000
MultiLoc	Höglund et. al., 2006	PrediSi	Hiller et. al., 2004
PSORT	Nakai and Kanehisa, 1991	WoLF PSORT	Horton et. al., 2007
SignalP	Nielsen et. al., 1997		

1.5.1. Computational Prediction Algorithms

Most of the prediction programs cluster proteins according to their functional domains. Before PROSITE (Sigrist *et al*., 2002, Sigrist *et al*., 2005, De Castro *et al*., 2006 and Sigrist *et al*., 2012), localization predictors were trying to catch localization associated patterns. The simplest method to find localization associated pattern is the alignment of the similar functioning proteins. As the conserved region of the alignment was the functional consensus sequence, the result of the prediction was actually predicting the function rather than the localization. Soon after the powerful tool, PROSITE, released for the annotation of the proteins, localization predictors started using PROSITE patterns, PROSITE profiles and PROSITE ProRules as the classification rules. UniProtKB/Swiss-Prot database is completely extended with PROSITE for protein annotations. The patterns, profiles or ProRules of

the PROSITE are avaible online for various demands at the site http://prosite.expasy.org/prosite.html . For example, PS00217 is used as a cytoplasmic membrane classification rule in PSORT. PS00217 is released in 2006 by PROSITE as a sugar transport protein signature.

PS00217 Statistics are listed below:

Consensus pattern:

[LIVMF]-x-G-[LIVMFA]-{V}-x-G-{KP}-x(7)-[LIFY]-x(2)-[EQ]-x(6)-[RK]

Sequences in UniProtKB/Swiss-Prot known to belong to this class: 370

- **detected** by PS00217: 229 (**true positives)**
- **undetected** by PS00217: 141 (138 **false negatives** and 3 'partials')

Other sequences in UniProtKB/Swiss-Prot detected by PS00217:

192 **false positives** and 2 unknowns.

Because of the fact that PS00217 annotation pattern bearing proteins are cytoplasmic membrane proteins, PSORT uses this sugar transport protein signature as a localization classifier. Predictors also uses NCBI CDD database, annotation linked domain IDs, as a localization classifier. As the basic methodology of the localization prediction is already explained, a brief information is given below about three different prediction programs that we used in our previous immunoproteomics study.

1.5.1.1 PSORT

PSORT is one of the most widely used computer programs to predict cellular protein localization. Amino acid sequence and its source origin are used as inputs and their analysis are conducted by several classification rules based on various sequence features of known protein sorting signals. As an output, it reports the possible localization of the input protein with respective percentages and applied classification rules (http://psort.hgc.jp). One of the five versions of this program, PSORTb v3.0.2, is the most precise bacterial localization prediction tool. PSORTb is used for all prokaryotes including

bacteria and archaea. In 2010, it was claimed that PSORTb v3 is the most precise prokaryotic SCL predictor (Yu *et. al*., 2010). It has both web server and open source standalone version. While sorting protein into the localization, this program uses methods and motifs such as CMSVM, cytoSVM, ECSVM, ModHMM, OMPMotif, OMSVM, PPSVM, Profile, SCL BLAST, SCL BLASTe and Signal Sequences for the decision. The secreted and surface associated motifs of PSORT are listed in Appendix C.

1.5.1.2. PrediSi

PrediSi (PREDIction of SIgnal peptides) is a web server based computer program to predict signal peptide sequences and their cleavage positions for bacterial and eukaryotic amino acid sequences (http://www.predisi.de/home.html). It was trained with the protein data of SwissProt database (Hiller *et al*., 2004). It performs calculations in real time with high accuracy such that 10 seconds is enough for 20000 eukaryotic sequences analysis. The program uses a position weight matrix improved by frequency correction considering amino acid bias.

1.5.1.3. SignalP

SignalP is a web based server program to predict the existence of transmembrane domains and the location of signal peptides in amino acid sequences. The program can be used for both prokaryotes and eukaryotes. A combination of several artificial neural networks is used to predict the signal peptide, its cleavage site and transmembrane regions (http://www.cbs.dtu.dk/services/SignalP/).

CHAPTER 2

MATERIALS AND METHODS

2.1. Databases Used in This Research

In this research we used several databases to construct our own datasets. NCBI PDB, Uniprot, AntigenDb, USPTO and PubMed databases were searched with the Geneious R8. All of the proteins in the protein databases are sorted with several unique and descriptive information. This information is embedded under a specific protein identifier (Entrez gi's). Every researcher can construct specific databases by retrieving targeted information from that gi's encoding structures. In order to show how descriptive data is embedded in a database, the Sequence 584 is given below as an example. The important information entries downloaded for the formation of immunogenic dataset are underlined. While searching the whole patent database, "immunogenic", "vaccine" and "epitope" are used as keywords.

LOCUS ADT43123 310 aa linear PAT 13-DEC-2010
DEFINITION Sequence 584 from patent US 7838010.
ACCESSION ADT43123
VERSION ADT43123.1 GI:314817799
DBSOURCE accession ADT43123.1
KEYWORDS .
SOURCE Unknown.

ORGANISM Unknown.
Unclassified.
REFERENCE 1 (residues 1 to 310)
AUTHORS Bensi,G., Grandi,G., Norais,N. and Ortega,M.J.R.
TITLE Immunogenic and therapeutic compositions for Streptococcus pyogenes
JOURNAL Patent: US 7838010-B2 584 23-NOV-2010;
Novartis Vaccines and Diagnostics S.R.L.; Siena;
IT;
REMARK CAMBIA Patent Lens: US 7838010
FEATURES Location/Qualifiers
source 1..310
/organism="unknown"
Region 8..301
/region_name="CorA"
/note="CorA-like Mg2+ transporter protein; pfam01544"
/db_xref="CDD:250695"
Region 14..303
/region_name="EcCorA_ZntB-like_u2"
/note="uncharacterized bacterial subfamily of the
Escherichia coli CorA-Salmonella typhimurium ZntB family;
cd12827"
/db_xref="CDD:213361"
Site order(15,48,63,65,81,96..99,101..102,127,130,179..180,183,
206)
/site_type="other"
/note="Cl binding site [ion binding]"
/db_xref="CDD:213361"

Site order (36,39..40,49,51,105,118,121,125,129,132..133,136,
139..140,142..143,146,158..159,161..162,165..166,169..170,
172..173,177,179..180,183..184,187..188,200..201,203..204,
207..208,210..211,214..215,217..218,220..222,225..232,
234..246,248..250,252..257,259..266,277..278,281,286,299,
302)
/site_type="other"
/note="oligomer interface [polypeptide binding]"
/db_xref="CDD:213361"

ORIGIN

1 mptlpkassa itlinldqlt ehdqealvse glidaevfdy akdknetsfm eenkekltiv

61 fqildrqees yhpnhiprvi pitlflndqs lyilghdhsl tlieevfpdl desrsprhlv

121 fqlltaftkq yvplmdeiaq qrdkliealr qkanktnles lanlqsgtvy ilmgskqneq

181 mlaelkdmpg qqdleedeae qlrdaiiear qlsnmcdlnt rvlkeisssy nnvlsnnlnn

241 nvtsltifsi gisiiamvts fygmnvklpf akvdsvwfwi vlttslvall imvvmywyvh

301 krnrnaskri

2.1.1 Data Retrieval Tool: Geneious R8

In this study, Geneious R8 was used as the data retrieval tool. Being able to communicate with a number of public databases hosted by the National Centre for Biotechnology Information (NCBI), as well as the UniProt database, Geneious is a powerfull data retrieval tool. NCBI, Uniprot, USPTO and PubMed are storehouses of molecular biology. One can search databases through Geneious by entering search terms. For advanced search, there is more option button. With advanced settings one can add more terms or more database by ”+” icon. The result window can be imported in various formats such as csv or fasta. The result window can also be filtered by various statistical approaches embedded as a function in Geneious. Geneious has helper icons as well; ”Any” of the fields (if only one of the fields needs to match), or ”All” of the fields (if all of the fields must match). For the immunogenic patented database construction, we used several icons written below.

+“Any” Bacteria
-Viruses
+“All” Patented Bacterial PDB

With respect to the desired data, the filtering or adding several terms can be made. After the data is created, sequential homology filtering can be done

with the same program. In our study, the sequential filtering was performed by using CD-HIT. Geneious was also used for CLUSTAL alignment of the pattern candidates of PSMS.

2.1.2. LC-MS Analysis

Before trypsin digestion 1 lane of 8cm length SDS page is divided 12 equal parts. Eymann et al. (2004) procedures of in-gel tryptic digestion as well as peptide elution for LC-MS/MS was performed. LC-MS/MS analysis done in was performed in Greifswald, Germany with the same procedure mentioned above. SEQUEST software used for protein identified. MS/MS –based peptide and protein validation tool was Scaffold 2.02.01 . Peptide and protein identifications were accepted if they could be established at greater than 99.5 % probablility and contained at least two identified peptides.

2.2. Relative abundance of surface proteins

In order to get proper set of proteins for pattern analysis with PSMS, the proteome data should be rearranged. Data rearrangement means filtering the proteins according to their relative abundances with respect to the abundance of the reference protein. In this work, the relative abundance data were taken from the work of Tefon *et al*. (2010).

Basically, from relative abundance differences of identified proteins by LC-MS/MS relative spectral counts (RSC) can be calculated using Equation 1.

$$RSC = (n1/t1 + n2/t2 + n3/t3)/3 \qquad \text{(Eq. 1)}$$

where n1, n2, and n3 stand for spectral counts for the protein in Sample 1, Sample 2, and Sample 3 (for 3 individual runs including technical and biological replicates) and t1, t2, and t3 are the sampling depths (total spectral

counts) where the spectral counts are taken as the number of spectral counts of a peptide with respect to discriminatory peptides.

2.3. Secretion System Data Construction

Geneious R8 was used for the secretion system data collection. After secretion system related proteins are collected, the duplicated sequences were removed with the same program. Geneious R8 is a very useful program for collecting and searching descriptive information of any sequences in databases. Dataset was extracted into a fasta file to filter proteins whose sequence homology is higher than 50%. The filtering was done by using CD-HIT program (Li &Godzik, 2006).

Table 2.1. Secretion system review articles for keyword selection for search agent of Geneious R8

Holland et al., 2005	Type 1 Secretion System
Nivaskumar & Francetic., 2013	Type 2 Secretion System
Barison et al., 2013 and Cornelis 2010	Type 3 Secretion System
Christie et al., 2014	Type 4 Secretion System
Henderson et al., 2004	Type 5 Secretion System
Durand et al., 2014	Type 6 Secretion System

While collecting proteins with Geneious R8, several secretion system review articles (Table 2.1) and the proteins listed there are taken as search agent keywords.

2.4. Immunogenic and Cytoplasmic Protein Data Construction

Geneious R8 was used for the immunoproteomic and cytoplasmic protein dataset construction. After proteins were collected, the duplicated sequences were removed with the same program. Dataset was extracted into a fasta file to filter proteins whose sequence homology is higher than 50%. The filtering was done by using CD-HIT program (Li &Godzik, 2006). For the immunogenic protein database construction, we used AntigenDB antigen database (Ansari et al., 2010) for vaccine development, immunogenic proteins of Uniprot (http://www.uniprot.org) and patented vaccine sequences (ftp.ebi.ac.uk/pub/databases/fastafiles/patent).

2.5. Dipeptide and Tripeptide Repeat Analysis

The relation between protein functions and their localization with respect to their short amino acid repeats was shown by several research groups (Cedano et. al., 1997; Cou, 2001; Bhasin and Raghava, 2004; Tantoso and Li, 2007). At the beginning, our aim was to find a locational pattern from the subproteomic data. However some of "short pattern candidates (SPC)" were not true patterns, instead they were dipeptide and tripeptide repeats. Because of this reason, dipeptide and tripeptide analysis function was added to the program. ,For a better explanation of this issue, some relevant partial sequences are given below:

>gi|7246030|gb|AAA25087.2| pullulanase, partial [*Klebsiella pneumoniae*]
MLRYTCNALFLGSLILLSGCDNSSSSSSSGSPDNPGNPDNQDVVVRL
PDVAVPGEAVMATANQAV

>gi|646771818|gb|KDR41838.1| hypothetical protein BG61_15225 [*Burkholderia glathei*]

MMKGKITAALAAVLLGAVTAAPAQTARPGDAASMVKPPSAAGASP
ARPDNPNNPDNMPMKRPAPPPNSDRMLHNSPASDA
IAR

>gi|727744008|ref|WP_033859631.1| hypothetical protein [*Staphylococcus aureus*]
MIGIHWGGVPNEFNGAVFINENVRNFLKQNIEDIHFANDDQPNNPDN
PDNPNNPDNPNNPDNPNNPDEPNNPDNPNNPDN
PDNGDNNNSDNPDAA

>gi|696671037|ref|WP_033143506.1| hypothetical protein, partial [*Blautia producta*]
MSYEWAVGEKVPDNPDNPDNPDNPDNPDNPDNPDNPDNPDNPDPD
NPDNPDNPDNPDNPDNPDPDNPDNPDNPDNPDNPDNPDNPDNPDNP
DNPDPDNPDNPDNPDNPNNPDNPDNPNNPDYPD
>gi|549638426|ref|WP_022542844.1| hypothetical protein [*Bifidobacterium animalis*]
MNYVVVLNHNDQPTDLNGVTAGKYGVLVSNGKTYLHPAVSDTTA
AAAVAAPVDTGKDVSKAGSTVQIAGGDAIAVPARSALLLGPTAVMS
EPDNPDNPNNPDQPDKPSGDNAHNGANNAANGANNASASHGGTAS
TGSSIAVIIAMVIVLLIAGSGMLLR

Apart from the pattern analysis of PSMS which requires a user input defining query coverage, there is no special requirement to be listed as dipeptide or tripeptide repeat because of the fact that the repeating unit number is not constant. Even among the the members of the same species, the number of the repeating unit may change. For instance, pullunases, gi|694067326 and gi|7246030, are the proteins of *Klebsiella pneumoniae* where the first protein has a hextuple repeating units and the latter has a triple repeating units.

2.6 *k-fold* cross-validation

In order to estimate how accurately a predictive model will perform in practice, we undertook five-fold cross validation tests. We divided our datasets into 5 random groups to validate the candidate patterns. In general, k-fold test starts with the randomly equal number of k number sub-data created from the original dataset. After this division, k-1 training dataset is selected versus 1 test dataset. This statistical method is used for producing single estimation.

2.7 Enabling Technologies

Intended to be simple, C# ,with its strong typing, imperative, declarative, functional, generic, object-oriented and component-oriented, become a modern multi-paradigm programming language. In this study we have used C# while creating PSMS. With respect to the Java with its excellent BioJava libararies, C# is not a commonly used bioinformatics programing languages. On the other hand, with its powerful error debugging and extensive code analysis tools, C# is favorable by non-computer science (non-CS) programmers. For example, a programmer does not need a fasta reader, protein or nucleotide sequence identifier or even RegEX optimization for biological arrays in Java with its BioJava libraries. For performing same functions a C# programmer needs to have fasta reader sniplet, biological sequence identifier and optimized REgEx for pattern analysis.

Let the user input sequence is

>gi|512722207|ref|WP_016480259.1| Immunogenic secreted protein [Streptococcus agalactiae]

MKKKILILGLLLPFLLPATLLILIAGNLSDDTTTSQNGQETSLTAKEVA
NKANISEERAKDVIQILNYQLSHEGFSLAGSARSLAVAERESGFDPK
AINTSGGVAGYFQWSGWSSSVNGNRWASAPQRALESQVELKLMST
ELSGAYANVKSKM…AVMSVKGGYGETLAQYGHVAFIEAVNKDGT
FLISECNYNHTQDKPHYQVLSPQSYYSFAIK

In C# a programmer needs to have a fasta reader code sniplet for understanting this fasta file is protein sequence and the sequence starts with M and lasts with K aminoacid. And also the program also need a short sniplet for understanting X as a meaning of any aminoacid inorder to find [IVALF]NXXGGXXG pattern, written in green. For performing same function nothing special required in BioJava libararies.

2.8 Epitope Prediction

We have used epitope prediction programs for our patterns that have observed in immunogenic proteins (http://tools.immuneepitope.org/tools/bcell/iedb_input). The epitope prediction data is used to understand whether our pattern has localization association or just epitope association.

CHAPTER 3

RESULTS AND DISCUSSION

3.1. Development of the PSMS Approach and Flow-chart

In order to improve cellular localization prediction programs, new patterns, motifs and profiles are needed. Program algorithms, decision rule algorithms and mathematical modelling of the predictors have been studied for almost 3 decades. Predictors now can only be drastically improved with new findings of protein pattern/motif/profile. In the present study, the aim for construction of PSMS is to find new localization markers according to their root of secretion systems. PSMS was developed for proteomics groups to co-publish localization markers with their sub-proteomic data. The simplicity of the user interface is shown in Figure 3.1 and the output is shown in Figure 3.2. User can identify new patterns from their sub-proteomic data or try to predict the secretion root of a specific protein. PSMS is weaker on general proteins but it has a better sensitivity for pathogenic bacterial secretion system-related proteins

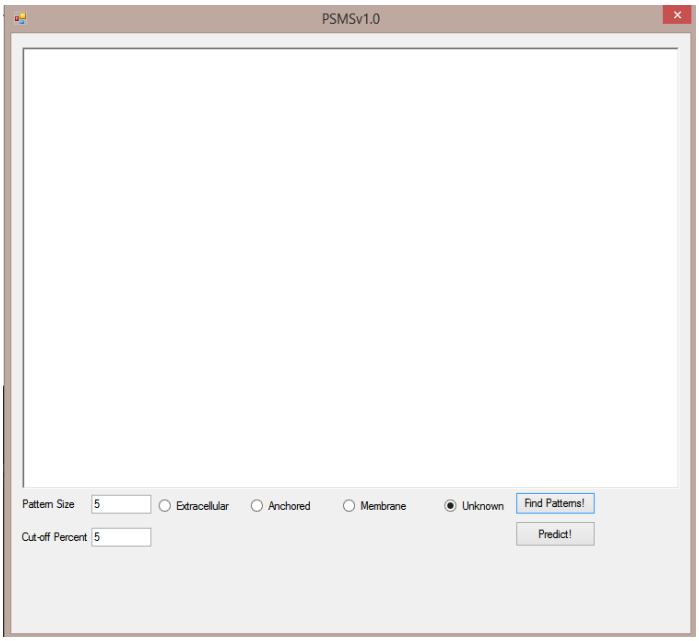


Figure 3.1. User interface of PSMS.

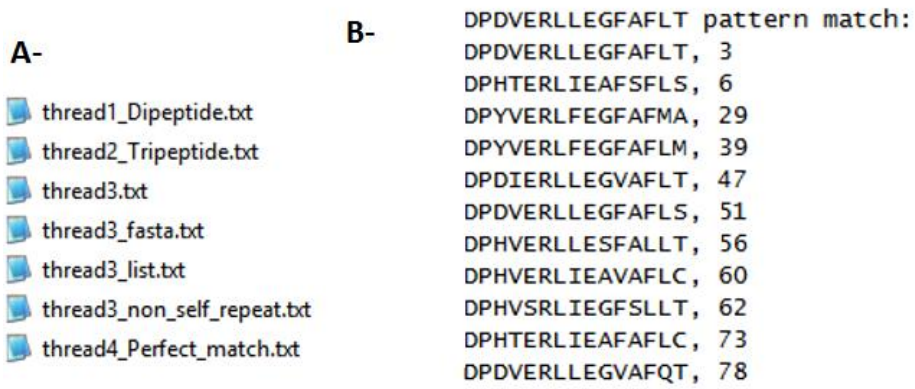


Figure 3.2. Results of PSMS. **A**. All of the 7 files created by using PSMS. **B**- Sample pattern analysis showing pattern and its matches where numbers are representing the protein numbers given by user input order.

3.1.1. Determination of Optimum Pattern Length for PSMS Analysis

Because of the deviation in distances between amino acids in 3D structure, pattern analysis should be used for small length amino acid sequences in order to prevent false positive results. Array length affects the deviation, so the highest number of amino acids in a predicted pattern was determined as 18 which was derived from the work by Katti et al. (2000). The lower limit was determined as 5, since one of the smallest localization motif "LPTXG" has 5 letters (Novick et al., 2000).

The allowed number of similar group of amino acids and the allowed number of different group of amino acids were derived from the analysis of short amino acid patterns in PSORTv3b motifs. The whole localization-associated motifs are presented in Appendix C It is known that PSORT holds the highest accuracy in protein localization (Yu et al., 2010). In order to determine the formula as a classification rule, graphs were plotted from the PSORT motif sets to see any regression of pattern length with allowed gap size and pattern length (Figure 3.3).
The optimum size of gaps for each pattern length is determined and the PSMS parameters are adjusted accordingly (Figure 3.3).

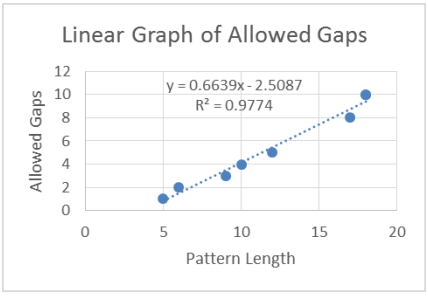


Figure 3.3. Formula derivation graphs for classification rule. Allowed gaps and pattern sizes were derived from PSORTb 3v.

There is a regression between a sequence length and the allowed number of different group amino acid. If there is no amino acid bias in user's sub-proteomic data, the allowed number of different group amino acids is derived from the graph in Figure 3.3. This classification threshold is smaller than the MotifSearch (http://motifsearch.com/). As the smallest number of the protein is 5 for the pattern search, we gave *A0A067AJ92_AMYMD, A0A081JCY9_9BURK, D3D1S5_9ACTO, E5U5N5_ALCXX, K9H531_9PROT* proteins for de-novo motif search. Even with a smaller dataset, MotifSearch gave 100 motifs. A higher number of motif candidates is not always preferred when all the candidates meant to be checked for registering as localization associated motif.

When a set of array is given to PSMS, each protein is divided into sub-arrays where each sub-array consists of k-mers with same length. All k-mers are later clustered with respect to the similar group of amino acids. This is done by giving every single amino acid a ground value with respect to its physicochemical similarity. A sample code snippet for doing this is presented in below.

```
public static string ReplaceNumbers(string protein)
    {
        protein = protein.Replace(" ", string.Empty);
        protein = protein.Replace('D', '1');
        protein = protein.Replace('E', '1');
        protein = protein.Replace('R', '2');
        protein = protein.Replace('H', '2');
        protein = protein.Replace('K', '2');
        protein = protein.Replace('S', '3');
        protein = protein.Replace('T', '3');
```

```
protein = protein.Replace('N', '3');
protein = protein.Replace('Q', '3');
protein = protein.Replace('C', '4');
protein = protein.Replace('U', '4');
protein = protein.Replace('G', '4');
protein = protein.Replace('P', '4');
protein = protein.Replace('A', '5');
protein = protein.Replace('Y', '5');
protein = protein.Replace('I', '5');
protein = protein.Replace('L', '5');
protein = protein.Replace('M', '5');
protein = protein.Replace('F', '5');
protein = protein.Replace('Y', '5');
protein = protein.Replace('W', '5');
protein = protein.Replace('V', '5');
return protein;
```

By giving ground values to each amino acid, PSMS is able to cluster k-mers with respect to their similarity.

Let

Array1 = $aa_1aa_2aa_3aa_4\ aa_5aa_6aa_7aa_8\ldots\ldots aa_n$

Then 5 membered arrays are :

$aa_1aa_2aa_3aa_4aa_5$

$aa_2aa_3aa_4aa_5aa_6$

$aa_3aa_4aa_5aa_6aa_7$

$aa_4aa_5aa_6aa_7aa_8$

.

.

.

$aa_{n-4}aa_{n-3}aa_{n-2}aa_{n-1}aa_n$

For example, [IVALF]NXXGGXXG pattern of Type 1 secretion system have 159 counts in TISS database. Few of the sub-sequence that represent this pattern are given in below for visual explanation:

-----I**N**AS**GG**IL**G**-
-----I**N**AR**GG**IL**G**-
-----I**N**AK**GG**IL**G**-
-----I**N**AK**GG**VL**G**-
-----I**N**AK**GG**VL**G**-
-----I**N**AS**GG**VL**G**-
-----I**N**AE**GG**VL**G**-
-----I**N**AA**GG**IL**G**-
-----I**N**AA**GG**VL**G**-
-----V**N**AA**GG**LL**G**-
-----I**N**DA**GG**ID**G**-

Where

*grounding values of the isoleucine, leucine and valine are the same

*lysine, serine, alanine and glutamic acid are of different grounding value,

So, PSMS is able to identify these sub-arrays as identical. All of the process used in PSMS development is summarized in Figure 3.4 as a flow-chart.

From pathway 1 to 2.3 in Figure 2.2, all of the steps are for “pattern analysis” and from pathway 3 to 3.1 are “secretion system decision” steps. Both pathways are defined with user input. Depending on the proteomic approach, the user can select various pre-determined proteomic localization for the amino acid bias. Default option for PSMS analysis is unknown which states there is no amino acid bias in the proteome pool. Cedano et al. (1997) showed that there is a relation between amino acid usage and localization. The amino acid means reported in this article for the 3 protein classes were used as a reference abundance with respect to amino acid bias detection. In the pathway 1.3, the program checks whether a user defines any specific pattern length for the analysis. The frequency of a pattern observed in a proteome pool is also

defined by user where the default query coverage is %10. The default pattern analysis is 5-amino acid long subsequences.

The formula defined in Figure 3.3 used pathway 1.4 where allowed gap is defined. Before giving pattern candidates, PSMS also checks whether it is a true pattern or just a piled up one with self-repeating units of dipeptides or tripeptides. The amino acid repeat analysis is explained in Section 2.4.1. True pattern candidates are clustered with CLUSTAL W. This clustering is used in web-logo projection where any 3 colored amino acid position is defined as X. All other possibilities, having less color than 3, are written in brackets. X can be used as any amino acid for bio java libraries, but lesser defined visual studio libraries need [A-Z] for the RegEx requirements. After RegEx is defined, the validation of the pattern is tested on various databases. Validated pattern candidates can also be used in pathway 3 (Figure 3.4) to find whether there is a specific secretion system association in the proteome pool submitted by user. By using RegEx, secretion associated patterns found in this thesis were searched on user input protein sequences.

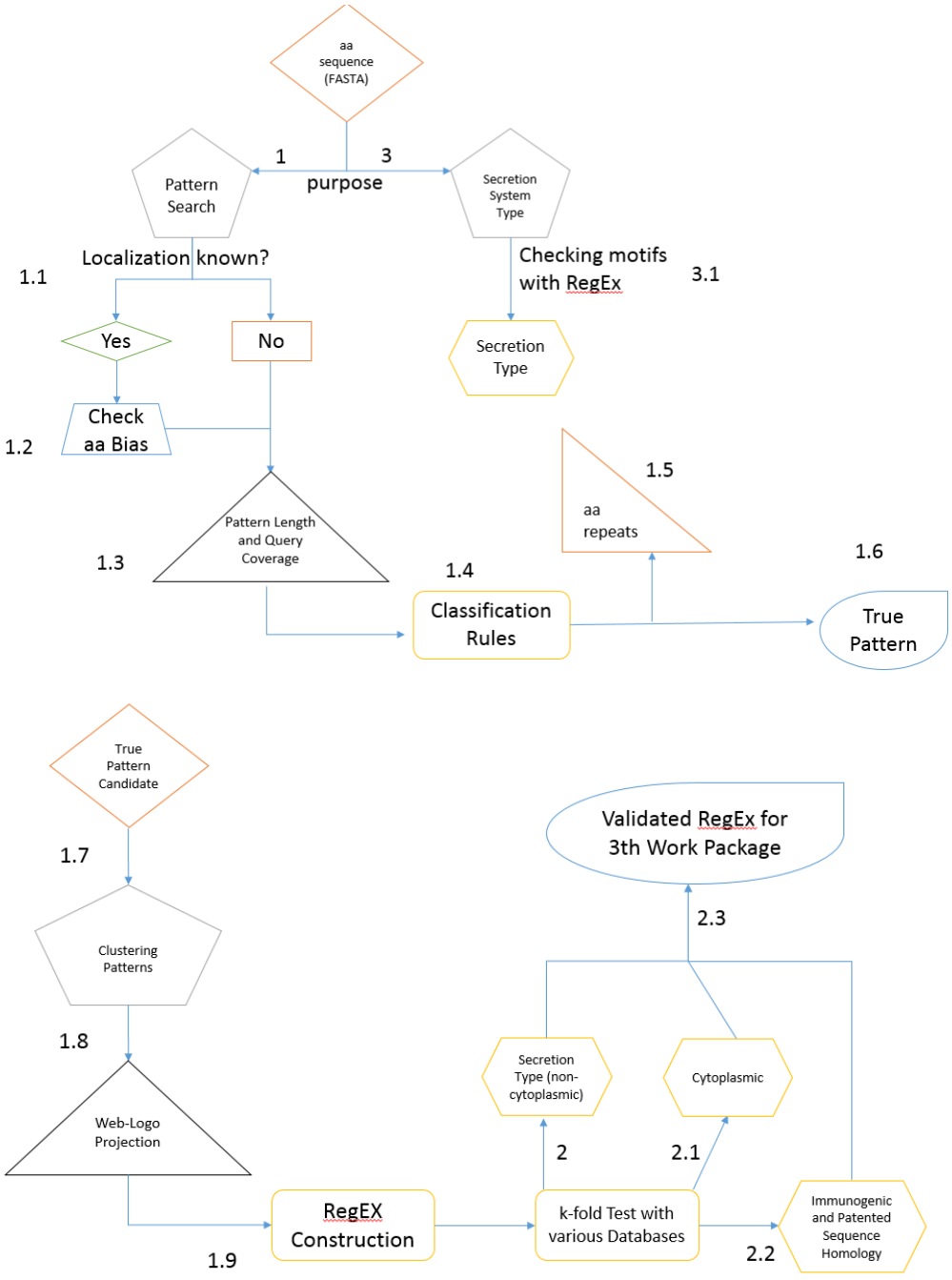


Figure 3.4. Flow-chart of PSMS development.

3.2 Database Construction

In this study we used several databases to construct our own datasets. We obtained 481 prokaryotic pathogenic sequences from AntigenDb then we searched the patented vaccine sequences from USPTO and gather 5166 additional pathogenic sequences. We applied a filtering step for proteins whose sequence homology is higher than 50% with the other sequences in the database. The filtering was done by using CD-HIT program (Li &Godzik, 2006) resulting in 1740 additional pathogenic proteins to AntigenDb database. For each of the pathogenic sequences we retrieved its secretion pathway information by using Geneious R8 based on their secretion type we divided our data into

6 different datasets, namely Type 1 secretion associated protein dataset (T1S), Type 2 secretion associated protein dataset (T2S), Type 3 secretion associated protein dataset (T3S), Type 4 secretion associated protein dataset (T4S), Type 5 secretion associated protein dataset (T5S), Type 6 secretion associated protein dataset (T6S). In order to validate our algorithm that it can truly distinguish secreted proteins we have established another database made up of 13308 truly cytoplasmic proteins that do not have immunogenic activity. After removal of homologous proteins by CD-HIT the number has decreased to 2582. Then to test the prediction ability of our algorithm in an independent data set and we have gathered 2429 immunogenic proteins form NCBI whose antigenic site information had not been determined previously.

Databases constructed and used throughout this study and the number of proteins in each are shown in Table 3.1

Table 3.1. Databases constructed and used in this study.

Database	**Number of proteins with CD-HIT 0.5 filtering**
TISS	954
T2SS	668
T3SS	381
T4SS	770
T5SS	221
T6SS	247
Secreted (orphan secreted protein database)	2533
Total Secretion System Related	5774
Cytoplasmic	2582
Immunogenic	2429

3.3 Pattern Analysis for *Bordetella pertussis* Secretome

High ratio of non-cytoplasmic proteins in the previous study of our research group led the team to examine LC-MS data with respect to the spectral counts and relative abundances (Tefon et. al., 2011 and 2013). The method described in Section 2.3.1 was used to create a new subset of LC-MS dataset where the proteins with concentrations higher than the BP0490 DNA polymerase III subunit beta were selected for pattern analysis. Pattern analysis of the rearranged dataset revealed more than 100 pattern candidates. The efficacy of the PSMS was tested by BLAST analysis with these candidate oligo-amino acid patterns or amino acid repeats. Batch BLASTp analysis of these candidate patterns found two of them novel: TLGLXGXGV and TXALAVAG. Rest of the patterns did not have localization specificity or they

were Prosite patterns. Any patterns matching with cytoplasmic proteins were filtered out from localization associated pattern list. We also found that several of the pattern candidates could be linked with Prosite functional domain patterns (membrane spanning parts, periplasmic proteins, and outer membrane attachment site profiles: PS51257).

A pattern is considered as insignificant/false result if it is not represented in proteins of several organisms. As a result of the study conducted, TLGLXGXGV was found as a candidate pattern. PSMS showed that this pattern is present in two completely different proteins (BP2922 and BP3831) in cytoplasmic protein filtered *Bordetella* surfacome dataset, and taxonomy BLAST result gave exact hits for more than 500 organisms (Figure 3.5A, Appendix D). Bacterial hits covered over 93% of total hits (Figure 3.1A) 61% of which belonged to gram-negative bacteria (Figure 3.5B). Ten percent of the gram-negative proteins were found to be cytoplasmic. (Figure 3.1C). The conservation of TLGLXGXGV pattern through evolution and the bias towards this pattern are clearly represented in Table 3.2 and Appendix D. The proteins that gave exact hits by BLASTp were also analyzed through PrediSi, Psort 3.0, SignalP and SVMtm. The majority of these hits were found to be non-cytoplasmic proteins in most organisms except for Archaea. The 56% of the archaeal proteins were found to be cytoplasmic proteins (Table 3.2.). The variable part of the TLGLXGXGV pattern was XGX.

It can be postulated that VGC, VGF, EGN or FGF amino acid combinations in the position of XGX are mostly found in cytoplasmic proteins, whereas LGX, RGX or AGA combinations are found in non-cytoplasmic ones (Table 3.2.). TLGLIGVGV sequence is found mainly in proteobacteria with ABC type-secreted proteins. The proteins that carry this sequence are periplasmic or membrane bound proteins. Wang et *al*., (2014) has recently claimed that GLIGV (as in TLGLIGVGV) sequence is rich in non-T4SS rather than T4SS.

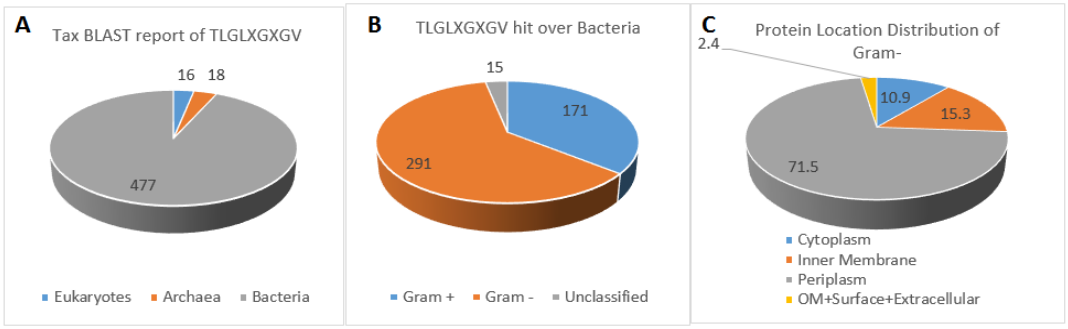


Figure 3.5. BLASTP analysis of the TLGLXGXGV pattern; A) Distribution of protein hits among Kingdoms, B) Distribution of proteins among bacteria, C) Distribution of Gram (-) protein hits with respect to cellular localization.

Another pattern that was scooped out from bulk possibilities is TXALAVAG which was derived from the BLAST results and represented more among Gram (+) bacteria. The pattern was detected in 483 organisms with 843 hits, 676 of which are bacterial (Appendix E). Surprisingly, 35% of the protein matches for this pattern were hypothetical proteins that were found by genome computational analysis (Appendix E). Similar to the TLGLXGXGV pattern, the type of mismatched amino acid in the TXALAVAG pattern is directly affected by the cellular localization of the protein. For example; in *Mycobacterium*, membrane proteins contain alanine whereas cytoplasmic proteins contain threonine in the position of X. As a result, it can be that apart from the TLGLXGXGV pattern which uses the mismatched amino acids with respect to taxa, X in TXALAVAG is rather determined by the cellular localization of proteins. The usage distribution of the mismatched amino acids are presented in Figure 3.5.

Table 3.2. BLASTP hit distribution of TLGL***XGX***GV pattern among prokaryotic organisms

Kingdom	Class	# of Org.	Total Hit	Amino acids of ***XGX*** in Pattern					
				UN	C	IM	P	S,EC	Location
Archaea	*Euryarchaeotes - Crenarchaeotes*	18	25	2	14	9	-	-	AGF,VGG are inner membranous
Bacteria	CFB group	28	37	1	4	32	-	-	All inner membranous ones have LGY
Bacteria	*γ-proteobacteria*	40	49	4	**9***	17	14	-	LGL and RGG are membranous and periplasmic
Bacteria	*α-proteobacteria*	21	25	4	6	12	-	3	TLG inner membranous
Bacteria	*β-proteobacteria*	181	241	-	-	12	229	-	IGX,RGX are periplasmic
Bacteria	*Δ-proteobacteria*	23	38	2	28	8	-	-	TGL,VGG,VGC cytoplasmic
Bacteria	*Planctomyces*	2	2	-	-	2	-	-	Insufficient data
Bacteria	*Cyanobacteria*	6	5	2	3	1	-	-	Insufficient data
Bacteria	*Firmicutes*	121	144	4	7	74	1	35	NGD surface
Bacteria	*Deinococcus-Thermus*	5	9	-	1	5	-	3	LGX inner membranous
Bacteria	*Actinobacteria*	47	62	2	15	40	-	5	LGX inner membranous
Bacteria	Unclassified bacteria	1	2		-	2	-	-	Insufficient data
Total		**509**	**674**	**26**	**90**	**215**	**244**	**47**	**LGX, RGX, AGA are non-cytoplasmic regardless of classification**

*New studies indicate that some of these peptidases are also secreted such as peptidase S8 which is classified as cytoplasmic protein by PSORT3b.

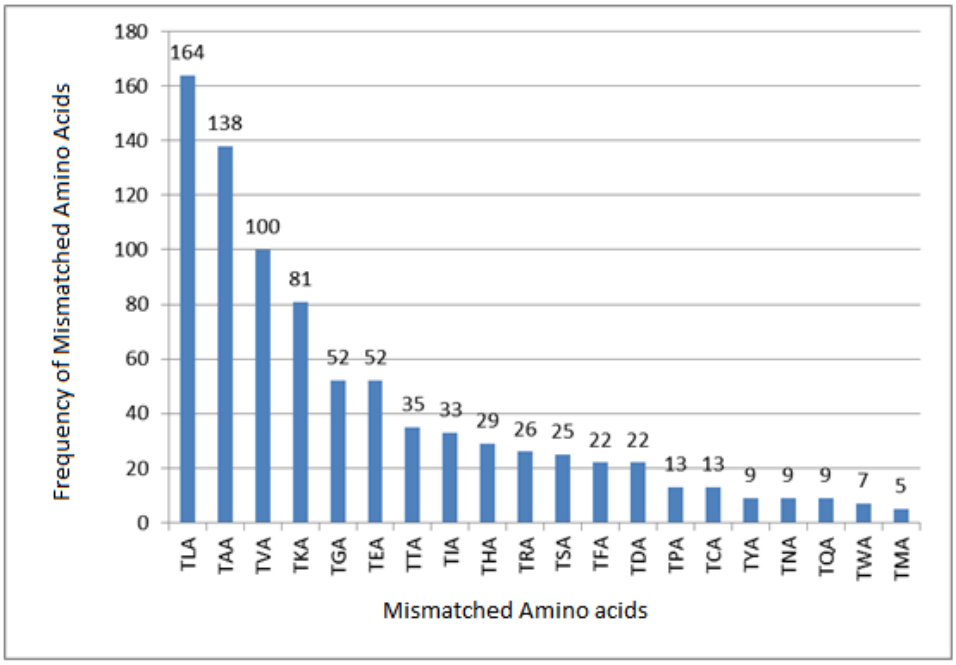


Figure 3.6 Mismatched amino acid frequencies for TXALAVAG pattern.

Of possible mismatched amino acids, leucine, valine, and serine were found generally in non-cytoplasmic proteins. The proteins carrying TLALAVAG, TVALAVAG and TSALAVAG are given in the Appendix E. The ABC type transport proteins and outer membrane porins may also represent this pattern, but this time isoleucine is replaced by a similar amino acid, valine.

3.4 Amino acid Repeat Analysis of *Bordetella pertussis* Secretome

In this study, the PNN repeat which was previously shown in the study of Katti *et al*. (2000) was improved as PXN where X can be N, D and S, but mostly D (Table 3.3). The exact hit table of the quadruple repeat is shown in Appendix F and functional domains carrying sequence abundance are in Table 3.4 and Table 3.5 PXN repeats were usually found on virulence factors located

on the surface of the pathogens, such as peptidases and collagenases. Although PSORTb prediction results show these proteases as cytoplasmic proteins, these virulence factors are located on the surface of the pathogen or they are secreted.

Table 3.3. The amino acid frequency of PXN repeat in Bacterial Uniref 50.

Aminoacid	Usage	Percentage	Aminoacid	Usage	Percentage
N	43,493	36.70%	I	1,519	1.30%
P	43,473	36.70%	L	1,306	1.10%
D	6,622	5.60%	Q	1,083	0.90%
T	3,349	2.80%	F	704	0.60%
G	3,328	2.80%	R	481	0.40%
S	2,842	2.40%	Y	472	0.40%
K	2,606	2.20%	H	188	0.20%
V	2,426	2.00%	C	164	0.10%
A	2,385	2.00%	M	131	0.10%
E	1,714	1.40%	W	120	0.10%

Table 3.4. Functions of triple and quadruple PXN repeat bearing domains which are found by BLASTp in NCBI non redundant prokaryotic database.

PXN repeat bearing domain functions	# of Hits
Surface binding and attachment proteins	686
Membrane spanning enzymes and attachment proteins	270
Outer membrane proteins	52
Surface linked proteases	322
Hypothetical surface associated proteins	3885
Inner membrane proteins and enzymes	209
Chemotaxis and motility proteins	69
Extracellulaar proteins	64
Transport machinary proteins	253
Pseudogenes (Repeat encoded genes)	29
Cytoplasmic proteins	148
Cytoplasmic binding proteins (DNA binding domains, regulatory proteins)	575
Proteins with no clues with current protein knowledge	3419

Table 3.5. PXN Hits in Uniref 50.

# of GI s captured from nonredundant prokaryotic protein database	15482
Total number of different proteins	9981
Approved noncytoplasmic proteins	5810
Approved cytoplasmic proteins	725
Proteins with no clues	3419

To find how frequent the rare repeats of our surfacome database are, a search was made in Uniref 50 database. We obtained 15482 hits on 9981 different proteins. About 7.3% of the proteins were found to be cytoplasmic whereas more than 58% of proteins were found to be non-cytoplasmic. For 3419 proteins bearing PXN repeat, any localization information has been associated with current protein knowledge.

3.5 Secretion System-associated Protein Patterns

It was observed that the mean pIs, MW and the amino acid usages of the proteins greatly differ in different types secretion systems (Table 3.6). pI, which is determined by amino acid frequency, is a strong characteristic for each type of secretion system. Therefore, random protein groups were formed for each secretion system type. For k-fold analysis, groups with mean pIs closer to the mean pI of the respective secretion system were used (Appendix G). The analysis resulted in 59 patterns related to six different secretion systems (Appendix L).

Table 3.6. Molecular weight, isoelectric point and sequence length statistics of each dataset.

	T1SS	T2SS	T3SS	T4SS	T5SS	T6SS	Cytoplasmic
Mean Length / STD	340 / 105.2	394 / 219.4	340 / 105.2	445.9 / 378.2	1246 / 793	448.6 / 268.2	383.8 / 370.3
MW (Mean)	36.852 kDa	43.385 kDa	33.221 kDa	49.166 kDa	128.982 kDa	50.149 kDa	41.292 kDa
IP (Mean)	6.79	7.4	6.75	7.36	5.88	6.47	6.79

The proteins used in this research are available with CDHIT 0.5 arranged at the site http://molmicrobio.metu.edu.tr/. The training dataset was analyzed with PSMS with %1, %5 and %10 query coverages, respectively. All of the pattern candidates (Appendix H) were clustered with CLUSTAL W to merge several patterns. After CLUSTAL alignment (Appendix I), the web–logos were plotted (Appendix K) to define RegEx formulations of the patterns (Appendix L). All pattern candidates were cross checked in test dataset. *k*-fold cross validity gave single estimations for the pattern candidates where their hit counts between *k-1* set and test dataset as a predictive model. The success of these predictive models of each pattern was next tested on cytoplasmic and immunogenic protein datasets. All of the immunogenic proteins were also analyzed by PSORT. Seventy five proteins that was shown by PSMS to be related with secretion systems were listed as unknown or cytoplasmic by PSORT. Although PSORT is known to make the most accurate predictions (Yu et al., 2010), the present study clearly showed that it gives misleading results for pathogenic proteins On closer inspection, failed predictions of the immunogenic dataset were found to mostly belong to intracellular pathogens of the following genera: *Neissera meningitides* (Facultative intracellular), *Mycobacterium tuberculosis* (Obligate intracellular), *Bordetella pertussis* and *Bordetella parapertussis* (Facultative intracellular), *Salmonella enterica* (Facultative intracellular), *Brucella abortus* (Facultative intracellular), *Yersinia pestis* (Facultative intracellular), *Streptococcus pyogenes* (Facultative intracellular), *Chlamydia trachomatis* (Obligate intracellular), *Francisella tularensis* (Facultative intracellular), *Mycobacterium leprae* (Obligate intracellular), *Listeria monocytogenes* (Facultative intracellular). The effector proteins of these pathogens were successfully predicted by using PSMS.

The proteins were randomly distributed into 5 groups by allowing mean pIs of each group closer to the mean pI of original protein dataset for five-fold test. The patterns of training datasets (4 groups) were searched on test datasets for hit counts to determine validity of each pattern candidate.

The patterns of training datasets were applied to test datasets. Although we started with 5 amino acids as the length for pattern search, CLUSTAL Alignments of the pattern candidates increased smallest pattern length to 9 amino acids (Appendix L). Clustering of the pattern candidates is crucial for handling huge pattern candidate list. For example, 133 different 5-aa-long pattern candidates aligned on top of each other to form the final 9-aa-long pattern [IVALF]NXXGGXXG:

```
5aa        INAAG
5aa          AAGGK
5aa              GGKLG
9aa        INAAGGKLG
```

The sensitivity of the pattern search with Regex algorithms is directly related with the pattern annotation. Web-logo projections were used to construct Regex form of the patterns. Based on physiochemical properties of amino acids, web-logo projection uses five different colors while plotting amino acid distribution of sequences. Observed amino acids in web-logo projection were written in brackets for Regex representations. Any amino acid position containing more than three colors were represented as X which indicates there is no physiochemical requirement for the 3-colored position in the pattern. Pattern performances were tested with immunoreactive dataset and results are given in Appendix O.

3.5.1 PSMS Analysis of TISS

Throughout this work we have used two different immunogenic data where one has fully understood and their epitope regions are known. And the other immunogenic database is constructed from PubMed where they are identified by immunoproteomic studies and their epitope regions are not fully studied so they are not found in descriptive information inside the NCBI PDB. We found 7 distinct TISS related patterns (Table L.1, Appendix L). The performance of the patterns were tested with our immunogenic dataset. Whole database was analysed with PSORTb version 3.0.2. The exact hits of each patterns in immunogenic database are given in Appendix O.

The P1 of TISS was represented 11 times in immunogenic database. The hit count decreased to 7 when the database was filtered by CDHIT for a sequence similarity less then 50%. Six of the 7 hits were patented (USPTO) vaccine sequences. The non-patented YP_008116936 sequence is an immunogenic protein of *Streptococcus agalactiae*. PSORTb could not localize this protein (categorized as unknown location). Another misleading result for P1 carrying proteins was ADT42608, a patented protein sequence (US 7838010). PSORTb localized this sequence as cytoplasmic membrane. The analysis of orthologues for all phosphate-binding protein pstS1 of *Streptococcus pyogenes* might cluster this protein as a pathogenicity-related secreted protein (http://www.xbase.ac.uk/genome/). Xbase is a server predicting pathogenicity proteins from their genomes. Because of the fact that many patented protein sequences for the vaccine technology are incomplete protein sequences focusing on epitopic regions, whole protein sequence or the orthologues sequence information should be used for localization prediction.

The probabilities of amino acids for each X block of P1 is shown in Figure 3.3. P1 was found in 117 different proteins in T1SS dataset (Appendix N, Table N.1). Three of four X blocks of P1 pattern was mainly represented by a single amino acid. The first two X blocks belong to alanine with a probability

over 50% whereas the fourth one was leucine with 47% probability. On the other hand, the third X block was represented by both valine (50%) and isoleucine (35%).

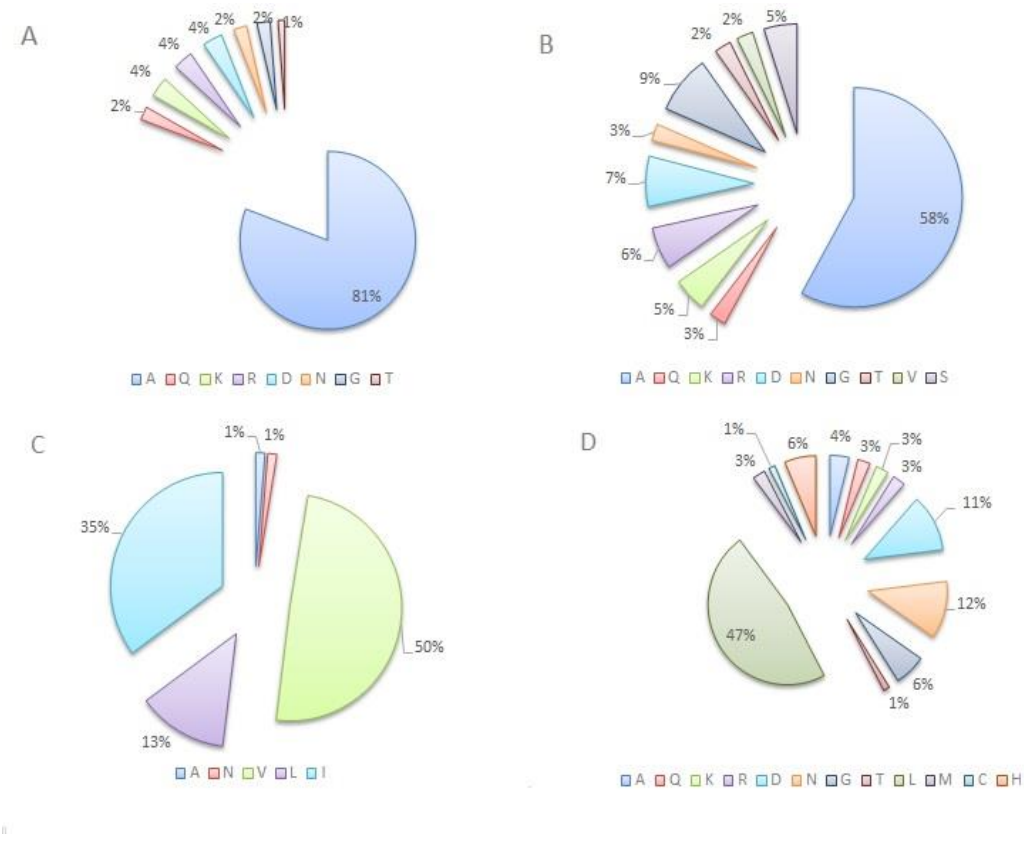


Figure 3.7. Amino acid percentages in every X blocks of 117 NXXGGXXG carrying proteins. **A**. Amino acid distribution on first X block of the pattern. **B**. Amino acid distribution on second X block of the pattern. **C**. Amino acid distribution on third X block of the pattern. **D**. Amino acid distribution on fourth X block of the pattern.

When we examine the pattern P1 of TISS, the X blocks of the pattern were not randomly distributed, but instead the majority of the blocks were formed with a hydrophobic amino acid alanine (Figure 3.3). In the literature, the nearest form of P1 was mentioned by Delepelaire (2004) as **GGXGXDXXX**

repeats. Although we found that P1 can be observed more than once in a protein, this was not obligatory. The X box with an aspartic acid (Figure 3.3. **D**) is valid for soil bacteria like Actinobacteria (*Streptomyces*) or nitrogen fixing ones (*Rhizobium*), but not valid for poly-beta-hydroxybutyrate-producing marine bacterium (*Oceanicola*) or a plant pathogen (*Agrobacterium*).

Like P1, P2 also existed in patented immunogenic protein dataset. Three patented sequences that are vaccine candidates of *Streptococcus pyogenes* beared P2. ADT42587 Sequence 47 from patent US 7838010 was one such protein that bears P2. Nineteen of the T1SS-related protein dataset was found to carry P2 in their sequences. A periplasmic protein of *Pseudomonas putida* (AFK70210.1), a transport mechanism protein *Clostridium botulinum* (WP_011948208.1) and a transport mechanism protein of an anaerobic extremophilic bacterium *Caldanaerobacter yonseiensis* (ERM91761.1) were among many examples that showed how P2 is conserved among various groups of bacteria. Although PSMS revealed that more than 10% of the immunogenic proteins are TISS-associated proteins, P3 pattern could not be detected among these proteins.

One of the highly counted pattern in immunogenic and patented protein dataset examined for TISS was P4 (253 raw hits /55 CDHIT 0.5). Surprisingly, P4 also gave hits for various prokaryotic EF-Tu on patented protein dataset (AJN58723 Sequence 269 from patent US 8889142). Although the house-keeping protein EF-Tu proteins are universally known as cytoplasmic proteins, there are various examples of immunoproteomic studies which reported the existence of this protein in secretome and outer membrane vesicles (OMVs, immunoreactive nature of this protein and possibilities of its usage as a novel immunogen against bacterial infections) (Kunze et al., 2004, Lopez et al., 2008, Nievs et at., 2010 and Lee et al., 2011). The protein needs to be directed to periplasm for packaging inside OMV for secretion. Not only proteins

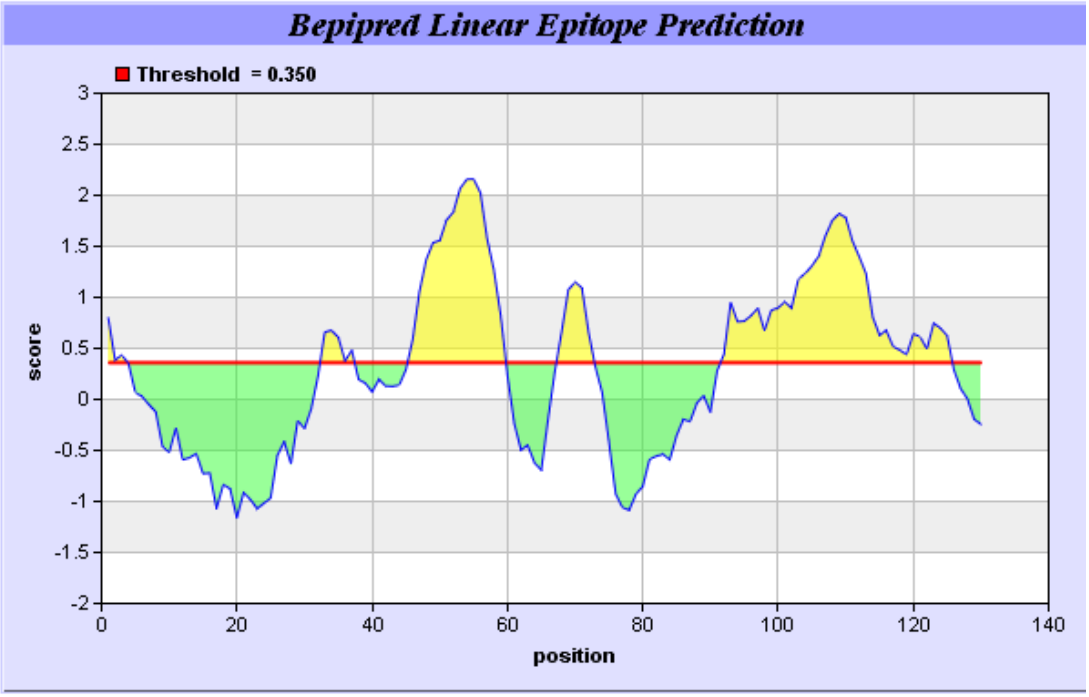
(periplasmic and outer membrane proteins) but also lipids, certain metabolites and even DNA can be packed heterogeneously inside OMV (Lee et al., 2008 and Kulp and Kuehn, 2010) and There are some other examples of cytoplasmic proteins leak into OMVs and as speculated by the authors, the concentration of house-keeping protein inside the cell might be high enough to account for this leak (Mashburn and Mercadé, 2005 and Kulp and Kuehn, 2010). Moreover, Caldas et al. (1998) showed that *E. coli* EF-Tu promotes functional folding and protein renaturation after stress, acting just like molecular chaperones. If EF-Tu is directed to periplasm for packing inside OMV, it needs a particular secretion system which might be specified by P4 recognized by T1SS. Epitopic region analysis of EF-Tu was made in order to distinguish its localization-associated pattern from its epitope pattern:.

>AJN58723 Sequence 269 from patent US 8889142
VTTESLETLVEQLSGLTVLELSQLKKLLEEKWDVTAAAPVVAVAGAA
AAGDAPASAEPTEFAVILEDVPSDKKIGVLKVVREVTGLALKEAKEM
TEGLPKTVKEKTSKSDAEDTVKKLQEAGAKAVAKGL

The amino acids in red represent the P4 region of the protein. The Antibody Epitope Prediction Tool which was mentioned in section 2.7 (http://tools.immuneepitope.org/
tools/bcell/iedb_input) was used for epitopic region prediction. Indeed, the third of the five predicted epitopic regions holds P4 region of the protein (Figure 3.4). To conclude, P4 is a both localization and epitope-related pattern and it is valuable for vaccine research in either way.

Sequence:

```
  1 VTTESLETLV EQLSGLTVLE LSQLKKLLEE KWDVTAAAPV VAVAGAAAAG DAPASAEPTE
 61 FAVILEDVPS DKKIGVLKVV REVTGLALKE AKEMTEGLPK TVKEKTSKSD AEDTVKKLQE
121 AGAKAVAKGL
```



Average:0.299 **Minimum:**-1.159 **Maximum:**2.159 **Threshold:** 0.350 Change

Click here to view plotted values in table format

Predicted epitopes:

No.	Start Position	End Position	Peptide	Peptide Length
1	1	3	VTT	3
2	33	37	DVTAA	5
3	46	59	AAAAGDAPASAEPT	14
4	68	72	VPSDK	5
5	92	125	KEMTEGLPKTVKEKTSKSDAEDTVKKLQEAGAKA	34

Figure 3.8. The epitope analysis of the protein AJN58723 with P4 region.

When we look at P5 of the TISS, the hit distribution of the pattern is mainly clustered on Lipoprotein 6 superfamily. One example was the immunogenic Variable small protein 24 of thick-born relapsing fever causing *Borrelia hermsii*. Members of this superfamily in pathogenic bacteria are known to play key roles like adhesion to host cells, modulation of inflammatory processes, and translocation of virulence factors into host cells, and shown to constitute potential vaccines. (Kovacs-Simon *et al.*, 2011).

Examination of P6-carrying proteins revealed that there is no bias for a certain protein superfamily. 10 kDa Chaperonin (*Helicobacter pylori*), immunogenic protein Mpt63 (*Mycobacterium* sp.), periplasmic immunogenic protein WP_032796870 (*Streptomyces* sp.), hypothetical protein WP_003911592 (*Mycobacterium tuberculosis*), immunogenic protein Bcsp31-2 (*Sulfitobacter* sp.), possible secreted protein CAC32006 (*Mycobacterium leprae*) and immunogenic patented (US8673316 sequence 10) of *Flavivirus* vaccine protein are few of the various immunogenic proteins that carry P6 in their sequences.

P7 could also be detected in various immunogenic protein families. When we compared the proteins of P6 and P7, we found that nearly half of the proteins of P7 have orthologues sequences with P6 proteins (21 over 47). Many of the rest of nonhomologous protein sequences were located at outer membrane or secreted ones.

3.5.2 PSMS Analysis of T2SS

Although PSMS patterns were easily detected in patented or immunogenic sequences, P1 of T2SS is observed only in several patented prepilin peptidase sequences like AJM00229 sequence (US patent 8741304). Various bacteria and even some viruses have this enzyme on their surface and its role in pathogenicity has been demonstrated by many studies. Pathogenic forms of *E.*

coli have even a second prepillin peptidase gene in their genome (Francetić et al., 1998).

Analysis of P2-carrying immunogenic and patented sequence dataset revealed that WP_022025931 immunogenic protein of *Clostridium sp.* CAG:75, outer membrane porin HofQ of *Chlamydia trachomatis* and potential drug and vaccine target UDP diphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase (ADL90375 Sequence 92 from patent US 7709009) (*Streptococcus pyogenes* and *Bacillus anthracis*) carried P2 region.

We could not detect any P3 in our immunogenic dataset, yet we found that P3 pattern for T2SS also gave hit with type 4 secretion system-associated proteins. Upon getting no hits on immunogenic dataset, we checked this pattern for cytoplasmic dataset to see if it is a pseudopattern or not. We could not observe any hit on cytoplasmic dataset, suggesting that this pattern is valid for non-immunogenic Type2 and Type 4 secretion-associated proteins with current protein database. Instead of checking all our databases we have checked for Uniref50.

Ten immunogenic proteins and three patented sequence counts were observed for P4 of T2SS. The proteins carrying P4 region had ATP-dependent fuctions. WP_029509212 cell division protein FtsK of *Leuconostoc lactis* has Pfam01580 ATPase domain and it is immunoreactive protein along with *Mycobacterium tuberculosis* (Jiang et al., 2014) and *Staphylococcus aureus* (Tedeshi et al., 2008) homologs. The immunogenicity of the proteins was detected via serological proteome analysis (SERPA) where patient sera crossed with pathogen soluble proteins. (Vytvytska et al., 2002) It was also shown that FtsK immunogenic proteins are linked with resistance and survivability of the *Staphylococcus aureus* (Vytvytska et al., 2012). Another ATP related domain carrying immunogenic patented protein was ADT43165 Sequence 626 from patent US 7838010 whose function is sugar transport in

Streptococcus pyogenes. This type of proteins are listed in various vaccine compositions (Garmony and Titball, 2004). Although the immunogenicity of P4 was validated with various immunogenic and patented proteins, it could not differentiate between Type 2 secretion-related proteins from Type 1 secretion-related ones. Half of the proteins carrying P4 of T2SS were Type 1 secretion-related proteins. The conserved positional existence of the P4 pattern led us checking from ExPASy Bioinformatics Resource Portal (www.prosite.expasy.org) if this pattern has a specific relation with functional domains. Altough the exact form of P4 is not listed in ExPASy server. P4 of our pattern is previously mentioned in the Prosite motif PDOC00017. ExPASy explanation of the domain also shows false positive results of the motif carrying proteins which are mainly Type 2 secretion associated proteins. Statistically validated Prosite motif which is carried by Type 1-associated proteins led us removing this pattern from T2SS. On the other hand, it must be noted that it is still an important pattern for non-cytoplasmic protein classification and search for proper vaccine candidates.

Another non immunogenic T2SS pattern, P5, was valid for T2SS- and T4SS-associated protein kinases. ExPASy analysis of P5- carrying proteins gave us PS00227 which is a GTP-binding protein. s. Unlike P4 form of ExPASy, P5 was more sensitive than ExPASy form as we have observed several cytoplasmic proteins whose carrying PS00227 domains whereas there is no cytoplasmic protein carrying P5. Like P5 of T2SS, P6 is also another nonimmunereactive T2SS associated protein pattern.

P7 and P8 were the other nonimmunogenic patterns. Although we observed that immunogenic YP_632115 protein of *Myxococcus xanthus* is carrying P7 pattern, the organism is a nonpathogenic saprophytic gram-negative bacterium. Because the big protein YP_632115 had more than 10 NCBI CDD hit and 8 different ExPASy motif matches, this could be due to cross

immunogenicity.. The rest of the proteins carrying P7 and P8 patterns were Type 2 and Type 4 secretion system-associated proteins.

P9 corresponded to several immunogenic outer membrane proteins like gi|440540208CCP65722 of *Chlamydia trachomatis*. The gi|899747306 chitinase protein of the same organism also carries a P9 region. Although many of the T2SS-dependent proteins are non-immunogenic, they can be potential drug targets by being pathogenicity-related hydrolytic enzymes like proteases, cellulases, pectinases, collagenases, phospholipases, and other virulence factors (Sandkvist, 2001)

3.5.3 PSMS Analysis of T3SS

A Total of 39 immunogenic counts were observed for the Type 3 associated patterns. Leo et al. (2012) postulated that T5SS and T3SS are evolutionary-related secretion systems. The P1 of the T3SS was a solid proof for this evolutionary relation. Immunogenic dataset analysis of the P1 gave hits to immunogenic T5SS related proteins like *vacA* gene product (VACA3_HELPX) of *Helicobacter pylori*. P2 of the T3SS is also immunogenic pattern with hit on iron and manganese responsive protein AJN59352 US patent 8889145 as a *Staphylococcus aureus* vaccine candidate. The patent consists of more than 16 different sequence version of the protein and all of them has a P2 region in their sequences. Epitope anaylsis of the protein showed that P2 itself is not an epitopic pattern.

P4 pattern held 26 immunogenic counts where 6 of them were patented sequences. CCP66010 polymorphic outer membrane protein repeat of *Chlamydia trachomatis*, CAR60209 outer membrane esterase of *Salmonella enterica*, 17kDa surface antigen of *Rickettsia*, ADT42758 of *Streptococcus pyogenes* Sequence 218 from patent US 7838010 and AFJ59155 surface antigen protein of *Pseudomonas fluorescens* are few examples of immunogenic proteins carrying P3 region in their sequences.

Immunogenicity of the P3, P5 and P6 patterns could not be detected in our 5166 immunogenic and patented sequences, but this does not mean that they are non-immunogenic patterns. In order to be placed in a database the information should be validified and meet certain statistics. In order to be listed in AntigenDB, the epitopic regions of the proteins should be validated via wet-lab. Immunoproteomic publications consisting of only immune-blot analysis might also give valuble informations but they need further anaylsis to be updated as immunogenic proteins in a database. Our immunoproteomics database is created with the proteins satisfying certain statistics to be counted as immunogenic proteins. When we increase the dataset with Uniref50 and look at the pattern hits in literature we can find more immunogenic proteins. For instance, AGH70349 nonstructural protein of porcine reproductive and respiratory syndrome virus is carrying our P3 region and it is an immunogenic protein. The protein is submitted by Veterinary Diagnostic Lab, China Animal Disease Control Center and the immunogenicity of the protein is published by Pei et al., 2009. This protein and its homologues forms could not be detected in either patented sequences or AntigenDB. The immunogenicity of th P5 carrying low calcium response locus protein D of intestinal pathogen *Vibrio parahaemolyticus* is not mentioned in databases yet homologous forms of the protein in *Yersinia pestis* and *Shigella flexneri* are found to be immunogenic (Andrews and Maurelli, 1992).

Another non-immunogenic pattern P6 has pilus assembly protein CpaC. Although in immunogenic database it is listed as immunoractive antigen, Forst et al, 1995 showed that this protein is immunogenic in *Neisseria gonorrhoeae*. Outer membrane protein PilQ, sequence 178 from patent US 7838010 and outer membrane secretin SsaC are examples of immunogenic proteins carrying P6 region in their sequence.

Sequence 42 and 43 from patent US 7709009, Sequence 399 from patent US 8889144, Sequence 4, 8, 10, 12 and 14 from patent US 8889145, seqeunce178 and 228 from patent US 7838010 and sequence 6, 10 and 12 from patent US 856873, outer membrane secretin of *Yersinia*. EHS39418, type 4 fimbrial biogenesis outer membrane protein *Pseudomonas aeruginosa* are patented and immunogenic examples of P7 and P8 carrying domains.

3.5.4 PSMS Analysis of T4SS

Pattern 13 from patent US 8703148, pattern 25 from patent US 8846342, Rhs family protein (Cell envelope biogenesis, outer membrane) of *Pseudomonas fluorescens*, pertussis toxin of *Bordetella bronchiseptica*, toxin TccC3 of *Yersinia pestis* are some of the P1-carrying vaccine component proteins. The mechanism of secretion of pertussis toxin via T4SS was reported by Hewlett and Donato (2007). Like P2 of T2SS, the P2 of T4 existed in pili and outer membrane porin proteins. Like the evolutionary relation between T3SS and T5SS as explained in Section 3.4.3, T2SS and T4SS show evolutionary relations.

P3-carrying proteins were the homologues proteins of P2-carrying ones. Immunogenic PS00875 Type 4 fimbrial biogenesis outer membrane protein of *Pseudomonas aeruginosa* carries P3 region in its sequence and the protectivity of vaccine formulation as a chimeric protein (type 4 pili+Endotoxin A) was reported earlier (Hertle *et al*., 2001).. Unlike truly T4SS-associated proteins, this protein was found by ExPASy as a T2SS-assosciated protein:

PS00875:
[GRH]-[DEQKG]-[STVM]-[LIVMA](3)-[GA]-G-[LIVMFY]-x(11)-
[LIVM]-P-

[LIVMFYWGS]-[LIVMF]-[GSAE]-x-[LIVMS]-P-[LIVMFYW]-[LIVMFYWS]-x(2,3)-[LV]-[FK]

Instead of C-terminal and mid-sequential patterns of T4SS captured by PSMS, PS00875 looked like to have N-terminal signals for T2SS. Also, this protein had only 32 posistive results in all UniRef 50.

Various homologous forms of PilA (which is a patented by US patent 8741304) hadve both P1 of T2ST2SS and P8 of T4SS pattern. In section 3.3.2, virulence association of peptidases was mentioned in Section 3.3.2. The higher homology of Prepilin peptidase and with PilA (between the aminoacids 10 to 125 from) validates their common pattern

P9 of T4SS is found in various immunogenic and secreted proteins. NP_719678 secreted VCBS domain protein *Shewanella oneidensis* MR-1, NP_437112 secreted calcium-binding protein *Sinorhizobium meliloti* 1021, YP_001976973 rhizobiocin/RTX toxin and hemolysin-type calcium binding protein *Rhizobium etli* CIAT 652, WP_037720566 large secreted protein, partial *Streptomyces sp.* CNQ329, YP_008580925 secreted protein *Leifsonia xyli* subsp. cynodontis DSM 46306, NP_938891 transport system secreted protein *Corynebacterium diphtheriae* NCTC 13129 were examples to the P9 carrying domains.

P10 of T4SS is one of the common pattern shared by T2SS. PilB and PilT of several bacteria carry this region. Understanding that the P10 carrying domains majorly found on motility associated proteins, ExPASy analysis conducted whether this pattern is associated with functional domains. ExPASy results showed that PS00662 having 26 bacterial hits for T2SS and 3 false positive hits. Three false positive results are actually T4SS secretion related proteins. Unlike PS00662, P10 has more pattern count for T4SS. The counts are mainly T4SS secretion-dependent motility.

3.5.5 PSMS Analysis of T5SS

Search for T5SS patterns indicated that that there are several patented and immunogenic proteins that carry T5SS association. 26 kDa periplasmic immunogenic protein, chimeric protein sequence of Staphylococcus *aureus* from patent US 8703148, Hemagglutinin of *Yersinia pestis* (Guedin et al., 2000), AJO27591 Sequence 135 of *Mycobacterium tuberculosis* from patent US 8486414, HtrA of *Haemophilus influenza* (protectivity reported by Loosmore et al., 1997), DegP-like immunogenic protein of *Chlamydia trachomatis* were among the immunogenic proteins that carry P1 in their sequence. P2, P3, P5, P6 and P8 corresponded to smaller numbers of immunoreactive proteins. AJO27566 sequence 97 of *Mycobacterium tuberculosis* complex is one of the patented immunoreactive protein carrying P2 (patent US 8486414). Outer membrane lipoprotein pcp of *Yersinia massiliensis* was carrying P3 seqeunce. YadA-like adhesion factor of *Rhizobium tropici* displayed P5 sequence. VirG, an outer membrane protein exposed to bacterial surface of *Shigella flexneri* had P6 sequence. YP_008863186 and MisL of *Salmonella enterica* had P8 sequence. P4 and P7 has many immunogenic proteins like P1. 17kDa surface antigen of *Sphingomonas*, cell surface protein of *Listeria monocytogenes*, outer membrane immunogenic protein of *Methylosinus trichosporium*, outer membrane protein PopD of *Pseudomonas aeruginosa*, OCEBH Immunogenic protein of poly-β-hydroxybutyrate-producing *Oceanicola granulosus*, immunogenic protein MPT63 of *Mycobacterium setense*, AGJ78007 immunogenic protein antigen 84 of *Propionibacterium avidum*, Sequence 256 from patent US 7838010 were those having P4 and P7.

3.5.6 PSMS Analysis of T6SS

We could not detect immunoreactive proteins in our immunogenic and patented database for T6SS.. Although human pathogens, including *Burkholderia mallei* and *B. pseudomallei*, *Vibrio cholera*, *Aeromonas hydrophila* and *Pseudomonas aeruginosa* have T6SS system in their genome,

the recent findings reveal that T6SS is used to target other bacteria and predators, efficiently killing or inhibiting competitors in natural environments (Hood et al., 2010, MacIntyr et al., 2010 and Shwarz et al., 2010).

Even though, many human proteobacterial pathogens contain T6SS-associated proteins, immune system does not show significant response to them since these proteins target other inhabited bacteria rather than host itself. This shows that immune system has a high perception of pathogenicity and in order to maintain a healthy microbiome, our immune system prefers to target only pathogenicity-associated proteins.

3.6 Comparison between PSORTb&PSMS

By filtering immunoreactive proteins and patented sequences, over five thousand proteins were reduced to 1740 proteins. Output file was first analysed by using PSORTb and 428 noncytoplasmic proteins were detected (130 outer membrane, 57 cell wall, 129 extracellular and 111 periplasmic). When the same output was analysed by using PSMS, 503 proteins were found to be related with secretion systems (TISS 277, T2SS 21, T3SS 67, T4SS 77 and T5SS 61 proteins). Many of the PSMS-localized proteins were the pathogenic effector proteins which are functional in host cells. Because of this reason, they do have cytoplasmic domains, for which PSORTb must have evaluated them as if they are cytoplasmic in the pathogen itself..

CHAPTER 4

CONCLUSION

4.1 Contribution

Missing potential vaccine candidates due to false localization predictions for massive proteomics data has been a general problem in vaccine discovery and development efforts. While trying to solve this problem in this study, we created the most comprehensive and inclusive database for secretion-associated proteins and immunogenic and patented vaccine proteins. Separate datasets for classification of the secretion system-associated proteins were constructed for the first time in this work and made available for the scientific community's use. (http://molmicrobio.metu.edu.tr/) Moreover, our immunogenicity-related database combined with CD-HIT filtered patented vaccine sequences as well as huge immunoproteomics data from our laboratory is four fold bigger than AntigenDb, the sole antigen database.

Additional contributions of our work is listed as below:

(1) The most pathogenic bacteria acquire resistance to current drugs. At the same time, there is an obvious need for the development and formulation of more immunogenic and much safer vaccines against deadly infectious diseases. A much better understanding and knowledge of the process of

pathogenesis will throw light on drug and vaccine discovery and development efforts. By focusing on pathogenicity related secretion system-associated proteins, we introduced a totally new practice for finding more surface localized, extracellular and immunogenic proteins by using constructed protein patterns. Our practice which is based on 43 distinct type-specific motifs of the virulence factors will greatly aid future efforts in the fields of molecular biology, biomedicine and pharmaceutical biotechnology

(2) Bacteria constitute a huge domain of prokaryotes. The uncountable variation in bacterial species is shown by several metagenomics studies. From this huge variation, the pathogenic bacteria becomes statistically unimportant for the generalization rules for protein localization predictors. PSORT program, the most widely used state of art computational tool for protein localization is quite powerful, however is not well suited for the prediction of pathogenic proteins. Owing to the selection of motifs in pathogenic sequences, the newly developed PSMS successfully identified potential vaccine candidates which were misleadingly localized as cytoplasmic and undeservedly excluded from vaccine research. PSMS program is now served to the scientific community as the most sensitive bioinformatic tool for immunoproteomics, reverse vaccinology, metagenomics and other molecular approaches to the pathogens.

(3) Analysis of the secretory system datasets by using PsMs revealed a the considerable structural homology, thus possible evolutionary relationships between the proteins of T2SS and T4SS for the first time. The different perspectives developed in this thesis were put in action, leading to important contributions to the literature.

4.2 Future Work

Currently, PSMS holds promise for pathogenic protein localization predictions and its patterns will find use soon. Although PSMS brought a new approach for identification of noncytoplasmic proteins by focusing on their root of secretion, its evolution and improvements will not solely involve localization predictions. In future, we aim at increasing the number of secretion-related patterns and the number of datasets while improving our algorithm with Hidden Markov Models. We plan to cluster secreted proteins in terms of their potential as vaccine candidates. Although we provide valuable information for vaccine researchers by predicting the root of secretion, not all secreted proteins are proper vaccine candidates. For testing the performance of PSMS, we created an improved immunogenic dataset improved which has even more members than AntigenDb In future. we plan to improve current dataset by considering potential protectivity of the proteins as well, hence converting PSMS to a vaccine candidate predictor.

REFERENCES

Andrade, M. A., O'Donoghue, S. I.& Rost, B. (1998). Adaptation of protein surfaces to subcellular location. JMol Biol, 276: 517-525.

Andrews, G.P., and Maurelli A.T. (1992) mxiA of Shigella flexneri 2a, which facilitates export of invasion plasmid antigens, encodes a homolog of the low-calcium-response protein, LcrD, of Yersinia pestis. Infect Immun., 60(8): 3287–3295.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res., 28: 45-48.

Bateman A, Murzin AG, Teichmann SA. (1998). Structure and distribution of pentapeptide repeats in bacteria. Protein Sci., 7: 1477–1480.

Bhasin M, Raghava G: ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic acids research

Bitter, W., et al. (2009). Systematic genetic nomenclature for type VII secretion systems. PLoS Pathog., 5: e1000507. 19876390

Blobel, G. (2000). Protein targeting (Nobel lecture). Chembiochem, 1: 86-102.

Boden, M. and Hawkins, J. (2005). Prediction of subcellular localization using sequence-biased recurrent networks. Bioinformatics, 21: 2279-2286.

Bork, P., et al. (1998). Predicting function:from genes to genomes and back. J Mol Biol, 283: 707-725.

Burkhardt J, Vonck J, Averhoff B. Structure and function of PilQ, a secretin of the DNA transporter from the thermophilic bacterium Thermus thermophilus HB27. The Journal of Biological Chemistry. 2011;286:9977–9984. doi: 10.1074/jbc.M110.212688.

Caldas, T.D., El Yaagoubi, A., Richarme, G. (1998). Chaperone Properties of Bacterial Elongation Factor EF-Tu*. JBio. Chem., 19(273):11478–11482.

Cedano, J., et al. (1997). Relation between amino acid composition and cellular location of proteins. J. Mol.

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

Congreve , M. , Murray , C. W. , Blundell , T. L. (2005). Structural biology and drug discovery. Drug Discov. Today, 10 : 895 – 907 .

Creighton, T.E., (1992). Proteins; Structure and Moleculer Properties. Macmillian Education. Chapter 1. 512.

Emanuelsson, O., et al. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol., 300: 1005-1016.

Feng, Z.P. (2002). An overview on predicting the subcellular location of a protein. In Silico Biol., 2: 291-303.

Fierer, N., Hamady, M., Lauber, C.L., Knight, R., (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc. Natl. Acad. Sci., 18:17994-17999.

Forest K.T., et al. (1996). Assembly and antigenicity of the Neisseria gonorrhoeae pilus mapped with antibodies. Infect Immun., 64(2): 644–652.

Francetić O., Lory, S., Pugsley, A.P. (1998). A second prepilin peptidase gene in Escherichia coli K-12. Mol Microbiol., 27(4):763-775.

Garmory, H.S., and Titball, R.W. (2004). ATP-binding cassette transporters are targets for the development of antibacterial vaccines and therapies. Infect. Immun., 72(12): 6757-6763.

Gibson D.L., White A.P., Rajotte C.M., Kay W.W. (2007). AgfC and AgfE facilitate extracellular thin aggregative fimbriae synthesis in Salmonella enteritidis. Microbiology., 153: 1131–1140.

Guda, C., et al. (2004). MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. Bioinformatics, 20: 1:785-794.

Guédin S., et al. (2000). Novel topological features of FhaC, the outer membrane transporter involved in the secretion of the Bordetella pertussis filamentous hemaglutinin. J Biol Chem., 29: 30202-30210.

Jack R.L., et al. (2004). Coordinating assembly and export of complex bacterial proteins. EMBO J. 23(20): 3962-3972.

Jiang Y., et al. (2014) Polymorphisms of FtsK/SpoIIIE protein in Mycobacterium tuberculosis complex strains may affect both protein function and host immune reaction. Int. J. Clinical and Exp. Med., 7(12): 5385-5393.

Hiller, K., et al. (2004). PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res., 32: 375-379.

Hsu, F., Schwarz, S., Mougous, J.D. (2009). TagR promotes PpkA-catalysed type VI secretion activation in Pseudomonas aeruginosa. Mol Microbiol., 72(5):1111-1125.

Katti, M. V., Sami-Subbu, R., Ranjekar, P. K., and Gupta, V. S. (2000). Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. Protein science: a publication of the Protein Society 9: 1203-1209.

Kulp, A., Kuehn, M.J.. (2010). Biological functions and biogenesis of secreted bacterial outer membrane vesicles. Annu. Rev. Microbiol., 64: 163-184.

Kumar, A. et al. (2002). Subcellular localization of the yeast proteome. Genes Dev, 16: 707-719.

Kunze, G., C. Zipfel, S. Robatzek, K. Niehaus, T. Boller, and G. Felix. 2004. The N terminus of bacterial elongation factor Tu elicits innate immunity in Arabidopsis plants. Plant Cell 16: 3496-3507.

Liao, L. and Noble W.S. (2003). Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. J. Comput. Biol., 10: 857–868.

Lee, K., et al. (2001). Identification and cloning of two immunogenic Clostridium perfringens proteins, elongation factor Tu (EF-Tu) and pyruvate:ferredoxin oxidoreductase (PFO) of C. perfringens. Res Vet Sci., 91: e80-86.

Lee, E.Y., Choi, D.S., Kwang-Pyo, K., Yong, S.G. 2008. Proteomics in Gram-negative bacterial outer membrane vesicles. Mass Spectrom., 27: 535–555.

Lopez JE, et al. (2008). High-throughput identification of T-lymphocyte antigens from Anaplasma marginale expressed using in vitro transcription and translation. J Immunol Methods, 332: 129–141.

Luthra, A., Mahmood, A., Arora, A., Ramachandran R. (2008). Characterization of Rv3868, an essential hypothetical protein of the ESX-1 secretion system in Mycobacterium tuberculosis. J. Biol. Chem., 283: 36532-36541.

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. (1999). A census of protein
repeats. J Mol Biol 293: 151–160.

Mashburn, L.M., Whiteley, M. (2005). Membrane vesicles traffic signals and facilitate group activities in a prokaryote. Nature. 15: 422-425.

Mott, R., et al. (2002). Predicting protein cellular localization using a domain projection method. GenomeRes., 12: 1168-1174.

Nair, R. & Rost, B. (2002). Sequence conserved for sub-cellular localization. Protein Science, 11: 2836-2847.

Nair R. and Rost B. (2003). Better prediction of subcellular localization by combining evolutionary and structural information. Proteins, 53: 917-930.

Nair, R. & Rost, B. (2002). Inferring sub-cellular localization through automated lexical analysis. Bioinformatics, 18: Suppl 1, S78-S86.

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

Nieves W, et al. (2010) Immunospecific Responses to Bacterial Elongation Factor Tu during Burkholderia Infection and Immunization. PLoS ONE 5(12): e14361.

Park, K.J. and Kanehisa, M. (2003). Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics, 19: 1656-1663.

Peia Y., et al. (2009) Porcine reproductive and respiratory syndrome virus as a vector: Immunogenicity of green fluorescent protein and porcine circovirus type 2 capsid expressed from dedicated subgenomic RNAs. Virology, 389: 91–99.

Petersen T. N., et al. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 8:785-786.

Pym, A.S., et al. (2002) Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines Mycobacterium bovis BCG and Mycobacterium microti. Mol. Microbiol., 46: 709–717.

Records, A.R. (2011). The type VI secretion system: a multipurpose delivery system with a phage-like machinery. Mol. Plant Microbe Interact., 24(7): 751-757.

Rost, B. (1999). Twilight zone of protein sequence alignments. Protein Eng., 12: 85-94.

Rondeau J. M. & Schreuder H., (2008). Protein Crystallography and Drug Discovery. The Practice of Medicinal Chemistry (Third Edition), 605–634.

Sawin, K.E. and Nurse, P. (1996). Identification of fission yeast nuclear markers using random polypeptide fusions with green fluorescent protein. Proc. Natl. Acad. Sci., 93: 15146–15151.

Schneewind, O., A. Fowler, and K. F. Faull. (1995). Structure of the cell wall anchor of surface proteins in Staphylococcus aureus. Science, 268: 103-106.

Schneider, G., and Wrede, P. (1994). The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. Biophys. J., 66: 335–344.

Sigrist C.J.A., et al. (2012). New and continuing developments at PROSITE. Nucleic Acids Res.
Sigrist C.J.A., et al. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform., 3:265-274.

Sutherland, H.G., et al. (2001). Large-scale identification of mammalian proteins localized to nuclear sub-compartments. Hum. Mol. Genet., 10: 1995–2011.

Szafron, D., et al. (2004). Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. Nucleic Acids Res., 32: 365-371.

Tedeschi G., et al. (2009). Serological proteome analysis of Staphylococcus aureus isolated from sub-clinical mastitis. Veterinary Microbiology., 134: 388–391.

Tefon, B. E., et al. (2011). A comprehensive analysis of Bordetella pertussis surface proteome and identification of new immunogenic proteins.

Wilkins, M., et al. (1996). Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. Biotechnol. Genet. Eng. Rev. 13: 19–50.

Vytvytska, O., et al. (2002). Identification of vaccine candidate antigens of Staphylococcus aureus by serological proteome analysis. Proteomics., American Society for Microbiology.

Yamaguchi H, Osaki T, Kai M, Taguchi H, Kamiya S. Immune Response against a Cross-Reactive Epitope on the Heat Shock Protein 60 Homologue of Helicobacter pylori. Burns DL, ed. Infection and Immunity. 2000;68(6):3448-3454.

Yu NY, et al. (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics, 26: 1608-1615.

Zhang, Z. and Wood, W.I. (2003). A profile hidden Markov model for signal peptides generated by HMMER. Bioinformatics, 19: 307-308.

http://psort.hgc.jp (Last accessed: 25.08.2015)

http://www.predisi.de/home.html (Last accessed: 26.08.2015)

http://www.cbs.dtu.dk/services/SignalP/ (Last accessed: 24.08.2015)

http://molmicrobio.metu.edu.tr/ (Last accessed: 25.07.2015)

APPENDIX A

SOFTWARE LICENCE

PSMS: Pathogenic Sequence Motif Search
Copyright (C) 2015 Orhan Özcan misalalemi@outlook.com

This program is free software: you can redistribute it and/or modify it under the terms of the GNU Affero General Public License as published by the Free Software Foundation, either version 3 of the License, or any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
GNU Affero General Public License for more details.

You should have received a copy of the GNU Affero General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.

APPENDIX B

Bordetella pertussis DATASET

Table B.1 *Bordetella pertussis* secreted proteins after data rearrangement

BAV0135	BB4237	BP0943	BP1605	BP2692	BP3659
BAV0939	BB4890	BP1054	BP1838	BP2747	BP3674
BAV1080	BB4940	BP1056	BP1852	BP2802	BP3732
BAV1088	BP0121	BP1071	BP1879	BP2818	BP3755
BAV1140	BP0205	BP1112	BP1881	BP2922	BP3783
BAV1159	BP0216	BP1119	BP1887	BP2953	BP3784
BAV1895	BP0345	BP1201	BP1900	BP2963	BP3819
BAV1960	BP0385	BP1277	BP2055	BP3080	BP3827
BAV2471	BP0454	BP1281	BP2067	BP3095	BP3831
BAV2755	BP0461	BP1285	BP2068	BP3159	BP3867
BAV2916	BP0479	BP1292	BP2072	BP3341	BPP1941
BAV3033	BP0558	BP1307	BP2219	BP3342	BPP2223
BAV3058	BP0561	BP1364	BP2256	BP3494	BPP3542
BAV3166	BP0562	BP1480	BP2315	BP3495	BPP3617
BB0324	BP0609	BP1487	BP2348	BP3551	BPP4198
BB3856	BP0698	BP1506	BP2352	BP3561	
BB3934	BP0782	BP1529	BP2396	BP3568	
BB3936	BP0840	BP1568	BP2418	BP3572	
BB4101	BP0856	BP1569	BP2497	BP3575	

APPENDIX C

PSORT CLASSIFICATION RULES

Table C.1 OMP rules used in PSORT3b

AAGAAG	AQAAVE	FGRSKD	GTLTVS	LGAATA	NNGTLI	SFLPSV
AAGKIS	AQTLEQ	FKLNYA	GTVSGL	LGALFR	NNNINA	SGLGRA
AALAAN	ARIEVG	FMGWMW	GVGINL	LGDIPV	NQLSVS	SGQTYN
AANANI	ASAREG	FRDFAE	GVKTDL	LGGDGI	NRSTLS	SGSFNF
AASAVE	ASNGLR.*LGRLGL	FSLKNS	GVLKTD	LGNLFK	NSIYID	SGSSSS
AASTTA	ATGAAV	FTGKGY	GYFDFR	LGRLGL	NTKTSS	SLAGTV
AAYRYS	ATLGLV	FVSLNA	HRIATL	LGTYLT	NTTINS	SLIALA
ADAADR	ATLTLT	GASAGV	IDNTST	LIACLS	NVTLQG	SLLAGS
ADLFPR	AVAVAL	GASSGY	IEARIV	LIDGKP	NYAAGG	SLLALS
AEIREK	AVDFHG	GDGGAI	IEQGTV	LLAATP	PGVSVG	SLLDVL
AELEQQ	AVDVAR	GDSLSD	IGAARA	LLDAQR	PLGLSD	SLLIGG
AETLAE	AVIAEV	GELSLS	IGRAGL	LLDVLD	PLLGDI	SLQQPL
AGAGAE	CFCLPL	GFIEDS	IGVLTD	LNLSIP	PTLDLT	SLSLPP
AGARYI	DGQDGD	GFNLNY	ISLTAN	LPIFTA	PVLAAD	SNITGG
AGGAIF	DGTLNL	GFSSRD	ISSPRL	LRPGMT	PVQVLA	SQLDWK
AGLAAL	DIQEFI	GGAISS	IYRNSP	LSAGVS	QANAAT	SRFSTS
AGLGAA	DIRVDG	GGAIYA	KEVLRD	LSERRA	QASWLA	SRLTLG
AGQASA	DNSKTD	GGANAA	KGGAIY	LSISGN	QFYLGA	SRPVAD
AGSGQV	DPRVKG	GGGAIY	KINEGP	LSLLPL	QGTVTL	SSSSSSS

Table C.2 Gram-negative motifs used in PSORT3b

Motif ID	Location	Motif
GGXGXD	EC	(GG.G.D.*){4}
PS01039	P	G[FYIL][DE][LIVMT][DE][LIVMF]...[LIVMA][VAGC]..[LIVMAGN]
PS01037	P	[GAP][LIVMFA][STAVDN]....[GSAV][LIVMFY]{2}Y[ND]...[LIVMF].[KNDE]
PS01040	P	[AG].{6,7}[DNEG]..[STAIVE][LIVMFYWA].[LIVMFY].[LIVM][KR][KRHDE][GDN][LIVMA][KNGSP][FW]
PS01157	P	GSYPSGHT
PS00635	P	[LIVMFY][APN].[DNS][KREQ]E[STR][LIVMAR].[FYWT].[NC][LIVM]..[LIVM]P[PAS]
PS00576	OM	[LIVMFY]..G..Y.F.K..[SN][STAV][LIVMFYW]V
PS00694	OM	(G[LIVMFY]N[LIVM]KYRYE)
PS00695	OM	([FYW]..G.GY[KR]F)
PS01151	OM	[VL][PASQ][PAS]G[PAD][FY].[LI][DNQSTAP][DNH][LIVMFY]
PS00875	OM	[GR][DEQKG][STVM][LIVMA]{3}[GA]G[LIVMFY].{11}[LIVM]P[LIVMFYWGS][LIVMF][GSAE].[LIVM]P[LIVMFYW]{2}..[LV]F
PS00834, PS00835	OM	((WTD.S.HP.T).*(AGYQE[ST]R[FYW]S[FYW][TN]A.GG[ST]Y))|((AGYQE[ST]R[FYW]S[FYW][TN]A.GG[ST]Y).*(WTD.S.HP.T))
PS01068	OM	[LIVMA].[GT].[TA][DA]..[DG][GSTP]..[LFYDE][NQS]..[LI][SG][QE][KRQE]RA..[LV]...[LIVMF].{4,5}[LIVM]....[LIVM]...[SG].G
PS00087	P	[GA][IMFAT]H[LIVF]H.{2}[GP][SDG].[STAGDE]
PS00123	P	[IV].DS[GAS][GASC][GAST][GA]T
PS00323	P	G[GN][SGA]G.R.[SGA]C.{2}[IV]
PS00401	P	K.[NQEK][GT]G[DQ].[LIVM].{3}QS
PS00538	CM	RTE[EQ]Q.{2}[SA][LIVM].[EQ]TAASMEQLTATV
PS00556	P	[LIVMA]{4}C[LIVMFA]T[LIVMA]{2}.{4}[LIVM].[RG].{2}L[CY]
PS00755	CM	[GST][LIVMF][LIVMFCA].[LIVMF][GSA][LIVM].P[LIVMFY]{2}.[AS][GSTQ][LIVMFAT]{3}[EQ][LIVMFA]{2}
PS00756	CM	[LIVMFYW]{2}.[DE].[LIVM][STDNQ].{2,3}[GK][LIVMF][GST][NST]G.[GST][LIV][LIVFP]
PS00757	P	NPK[ST]SG.AR
PS00968	P	[LIVFAG].[GASV][LIVFA].[IV]H.{3}[LIVM][GSTAE][STANH].{1,3}[STN]W[LIVMFYW]
PS00969	P	[EQ].{4}H.{5}[GSTA].{3}[FY].{3}[AG].{2}[AV]H.{7}P

Table C.2 Continued..

Motif ID	Location	Motif
PS00192	CM	[DENQ]...G[FYWMQ].[LIVMF]R..H
PS00193	CM	P[DE]W[FY][LFY]{2}
PS01303	CM	[GSDN]WT[LIVM].[FY]W.WW
PS00449	CM	[STAGN].[STAG][LIVMF]RL.[SAGV]N[LIVMT]
PS00713	CM	P.{0,1}G[DE].[LIVMF]{2}.[LIVM]{2}[KREQ][LIVM]{3}.P
PS00714	CM	P.G.[STA].[NT][LIVMC]DG[STAN].[LIVM][FY].{2}[LIVM].{2}[LIVM][FY][LI][SA]Q
PS00217	CM	[LIVMF].G[LIVMFA]..G.{8}[LIFY]..[EQ].{6}[RK]
PS00218	CM	[STAGC]G[PAG].{2,3}[LIVMFYWA]{2}.[LIVMFYW].[LIVMFWSTAGC]{2}[STAGC]...[LIVMFYWT].[LIVMST]...[LIVMCTA][GA]E.{5}[PSAL]
PS00274	EC	[KT]..NW..T[DN]T
PS00943	CM	N...[DEH]..[LIMF]D..[VM].R[ST]..R.{4}G
PS00077	CM	[YWG][LIVFYWTA]{2}[VGS]H[LNP].V.{44,47}HH
PS00221	CM	[HNQA].NP[STA][LIVMF][ST][LIVMF][GSTAFY]
PS00428	CM	[NV].{5}[GTR][LIVMA].P[PTLIVM].G[LIVM]...[LIVMFW][LIVMFW]S[YSA]GG[STN][SA]
PS00994	CM	R[LIVM][GSA]EV[GSA]ARF[STAIV]LD[GSA][LM]PGKQM[GSA]ID[GSA][DA]
PS00896	CM	G[LIVM]{2}.D[RK]LGL[RK]{2}.[LIVM]{2}W
PS00897	CM	P.[LIVMF]{2}NR[LIVM]G.KN[STA][LIVM]{3}
PS00942	CM	[QEK][RF]G.{3}[GSA][LIVF][WL][NS].[SA][HM]N[LIV][GA]G
PS01307	CM	A[LMF].[GAT]T[LIVMF].G.[LIVMF].{7}P
PS00594	CM	IG[GA]GM[LF][SA].P.{3}[SA]G.{2}F
PS00330	EC	D.[LI].{4}G.D.[LI].GG.{3}D

Table C.3 Gram-positive motifs used in PSORT3b

Motif ID	Location	Motif
PS00077	CM	[YWG][LIVFYWTA]{2}[VGS]H[LNP].V.{44,47}HH
PS00192	CM	[DENQ]...G[FYWMQ].[LIVMF]R..H
PS00193	CM	P[DE]W[FY][LFY]{2}
PS00216	CM	[LIVMSTAG][LIVMFSAG]..[LIVMSA][DE].[LIVMFYWA]GR[RK].{4,6}[GSTA]
PS00217	CM	[LIVMF].G[LIVMFA]..G.{8}[LIFY]..[EQ].{6}[RK]
PS00218	CM	[STAGC]G[PAG].{2,3}[LIVMFYWA]{2}.[LIVMFYW].[LIVMFWSTAGC]{2}[STAGC]...[LIVMFYWT].[LIVMST]...[LIVMCTA][GA]E.{5}[PSAL]
PS00274	EC	[KT]..NW..T[DN]T
PS00277	EC	YGG[LIV]T.{4}N
PS00278	EC	K..[LIVF].{4}[LIVF]D...R..L.{5}[LIV]Y
PS00369	C	G[LIVM]H[STAV]R[PAS][GSTA][STAMVN]
PS00589	C	[GSTADE][KREQSTIV].{4}[KRDN]S[LIVMF]{2}.[LIVM]..[LIVM][GADE]
PS00429	EC	ARP...K.S.TNAYNVTT..[DN]G...YG
PS00449	CM	[STAGN].[STAG][LIVMF]RL.[SAGV]N[LIVMT]
PS00713	CM	P.{0,1}G[DE].[LIVMF]{2}.[LIVM]{2}[KREQ][LIVM]{3}.P
PS00714	CM	P.G.[STA].[NT][LIVMC]DG[STAN].[LIVM][FY]..[LIVM]..[LIVM][FY][LI][SA]Q
PS00872	CM	[DG]...G...[DN].{6,8}[GA][KRHQ][FSA][KR][PT][FYW][LIVMWQ][LIV].[GAFV][GSTA]
PS00943	CM	N...[DEH]..[LIMF]D..[VM].R[ST]..R.{4}G
PS01022	CM	[GA][GAS][LIVMFYWA][LIVM][GAS]D.[LIVMFYWT][LIVMFYW]G...[TAV][IV]...[GSTAV].[LIVMF]...[GA]
PS01023	CM	[FYT]..[LMFY][FYV][LIVMFYWA].[IVG]N[LIVMAG]G[GSA][LIMF]
PS01072	CW	[LVFYT].[DA].{2,5}[DNGSATPHY][FYWPDA].{4}[LIV]..[GTALV].{4,6}[LIVFYC]..G.[PGSTA].{2,3}[MFYA].[PGAV].{3,10}[LIVMA][STKR][RY].[EQ].[STALIVM]
PS01219	CM	D[FYWS]AG[GSC].{2}[IV].{3}[SAG]{2}.{2}[SAG][LIVMF].{3}[LIVMFYWA]{2}.[GK].R

APPENDIX D

TAXONOMY BLAST OF TLGLXGXGV

Table D. Taxonomy of TLGLXGXGV

Taxonomy Report

root ..... 2554 hits 1667 orgs

. cellular organisms ..... 2548 hits 1663 orgs

. . Archaea ..... 55 hits 42 orgs

. . . Euryarchaeota ..... 50 hits 39 orgs

. . . Thermosphaera aggregans ..... 3 hits 2 orgs [Crenarchaeota;

. . . uncultured archaeon A07HB70 ..... 2 hits 1 orgs [environmental samples]

. . Bacteria ..... 2225 hits 1522 orgs

. . . Bacteroidetes ..... 64 hits 49 orgs [Bacteroidetes/Chlorobi group]

. . . Proteobacteria ..... 858 hits 552 orgs

. . . Firmicutes ..... 1030 hits 731 orgs

. . . Planctomycetaceae ..... 9 hits 6 orgs [Planctomycetes;

. . . unclassified Bacteria ..... 3 hits 3 orgs

. . . Acidobacteria ..... 6 hits 4 orgs [Fibrobacteres/Acidobacteria group]

Table D. Continued..

. . . Spirochaetales ..... 9 hits 8 orgs [Spirochaetes; Spirochaetia]
. . . Dehalococcoides mccartyi ..... 6 hits 3 orgs [Chloroflexi;
. . . Mycoplasma ovis ..... 3 hits 2 orgs [Tenericutes; Mollicutes;
. . Eukaryota ..... 268 hits 99 orgs
. . . Opisthokonta ..... 98 hits 64 orgs
. . . . Fungi ..... 55 hits 41 orgs
. . . Emiliania huxleyi CCMP1516 ..... 2 hits 1 orgs [Haptophyceae; Isochrysidales; Noelaerhabdaceae; . Viruses ..... 6 hits 4 orgs
. . dsDNA viruses, no RNA stage ..... 5 hits 3 orgs
. . . Sulfolobus turreted icosahedral virus ..... 2 hits 1 orgs [Rudiviridae; Rudivirus; unclassified Rudivirus]
. . . Yersinia phage YpsP-G ..... 1 hits 1 orgs [Caudovirales;
. . Rabbit hemorrhagic disease virus ..... 1 hits 1 orgs [ssRNA viruses; ssRNA positive-strand viruses, no

APPENDIX E

EXACT HITS OF TXALAVAG

ref|WP_031352282.1| membrane protein [Mycobacterium avium]
ref|WP_019306081.1| hypothetical protein [Mycobacterium avium]
ref|WP_044100456.1| membrane protein [Mycobacterium bovis]
gb|AHM72892.1| ChiP-III [Yersinia enterocolitica LC20]
ref|WP_044082017.1| membrane protein [Mycobacterium tuberculo...
ref|WP_003906384.1| hypothetical protein [Mycobacterium tuber...
ref|XP_004928840.1| PREDICTED: cytochrome b-c1 complex subuni...
ref|WP_018985861.1| hypothetical protein [Methylophilus methy...
ref|WP_007494848.1| hypothetical protein [Streptomyces zincir...
ref|WP_009587146.1| hypothetical protein [Acinetobacter sp. W...
ref|WP_019618652.1| hypothetical protein [Pseudoclavibacter f...
ref|XP_744839.1| hypothetical protein [Plasmodium chabaudi ch...
ref|WP_043941329.1| hypothetical protein [Weissella cibaria]
ref|WP_027936967.1| copper chaperone CopZ [Anaeroarcus burkin...
ref|WP_018704368.1| hypothetical protein [Anaeromusa acidamin...
ref|WP_035485767.1| hypothetical protein [Alicyclobacillus co...
ref|WP_029399125.1| MULTISPECIES: membrane protein [Mycobacte...
ref|WP_037180446.1| hypothetical protein [Rhodococcus fascians]
ref|WP_035836098.1| hypothetical protein [Cryobacterium roopk...

All of the rest organisms are given in CD version of the thesis

APPENDIX F

PROTEINS THAT HAVE QUADRUPLE REPEATS OF PXN

gb|EUZ23118.1| hypothetical protein O536_02847 [Staphylococcu...
gb|KCY43501.1| hypothetical protein J705_2842 [Acinetobacter ...
gb|EVM02052.1| hypothetical protein O902_01496 [Staphylococcu...
gb|EXC35763.1| hypothetical protein J455_4259 [Acinetobacter ...
gb|EZR77264.1| hypothetical protein W784_01618 [Staphylococcu...
gb|EUU81352.1| hypothetical protein O349_01868 [Staphylococcu...
gb|KCY05325.1| hypothetical protein J526_4165 [Acinetobacter ...
gb|EXC41991.1| hypothetical protein J455_2503 [Acinetobacter ...
gb|EXE32068.1| hypothetical protein J571_4155 [Acinetobacter ...
gb|KGW84967.1| hypothetical protein Y048_1641 [Burkholderia p...
gb|ADG60473.1| conserved hypothetical protein [Moraxella cata...
gb|EXD74565.1| hypothetical protein J488_1842 [Acinetobacter ...
gb|EXC21335.1| hypothetical protein J549_2607 [Acinetobacter ...
gb|EED65263.1| conserved hypothetical protein [Comamonas test...
ref|WP_040357625.1| chitin-binding protein [Corynebacterium c...
gb|EXV61936.1| hypothetical protein J835_3122 [Acinetobacter ...
gb|EKL51244.1| hypothetical protein ACIN5180_0001 [Acinetobac...
ref|WP_043375463.1| hypothetical protein [Comamonas testoster...
gb|EXS11313.1| hypothetical protein J654_2404 [Acinetobacter ...
ref|WP_031945428.1| hypothetical protein [Acinetobacter bauma...
gb|KGS04046.1| hypothetical protein X977_4078 [Burkholderia p...

All of the rest organisms are given in CD version of the thesis

APPENDIX G

DATASET VALIDITY

Table G.1 Amino acid statistics for grouped TISS proteins for k- fold cross validation

Group 1		Group 2		Group 3		Group 4		Group 5		Total	
ength: 726		Length: 815		Length: 1,252		Length: 752		Length: 722		Length: 1,252	
Sequences: 191		Sequences: 191		Sequences: 191		Sequences: 191		Sequences: 190		Sequences: 954	
Lengths of 191 sequences:		Lengths of 191 sequences:		Lengths of 191 sequences:		Lengths of 191 sequences:		Lengths of 190 sequences:		Lengths of 954 sequences:	
Mean: 341.9 Std Dev. 86.7		Mean: 332.8 Std Dev. 103.9		Mean: 346.5 Std Dev. 130.1		Mean: 343.5 Std Dev. 99.1		Mean: 335.3 Std Dev. 100.7		Mean: 340.0 Std Dev. 105.2	
Minimum: 69 Maximum: 726		Minimum: 47 Maximum: 815		Minimum: 53 Maximum: 1252		Minimum: 60 Maximum: 752		Minimum: 64 Maximum: 722		Minimum: 47 Maximum: 1252	
Molecular weight (mean): 36.936 kDa		Molecular weight (mean): 36.123 kDa		Molecular weight (mean): 37.506 kDa		Molecular weight (mean): 37.251 kDa		Molecular weight (mean): 36.444 kDa		Molecular weight (mean): 36.852 kDa	
Isoelectric point (mean): 6.74		Isoelectric point (mean): 6.77		Isoelectric point (mean): 6.95		Isoelectric point (mean): 6.83		Isoelectric point (mean): 6.69		Isoelectric point (mean): 6.79	
Extinction Coefficient (mean): 35,686		Extinction Coefficient (mean): 36,757		Extinction Coefficient (mean): 36,132		Extinction Coefficient (mean): 37,751		Extinction Coefficient (mean): 38,933		Extinction Coefficient (mean): 37,050	
Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
A: 7,612	11.7%	A: 7,168	11.3%	A: 7,661	11.6%	A: 7,280	11.1%	A: 7,150	11.2%	A: 36,871	11.4%
C: 466	0.7%	C: 431	0.7%	C: 439	0.7%	C: 464	0.7%	C: 393	0.6%	C: 2,193	0.7%
D: 3,967	6.1%	D: 4,040	6.4%	D: 4,089	6.2%	D: 4,055	6.2%	D: 3,863	6.1%	D: 20,014	6.2%
E: 3,595	5.5%	E: 3,452	5.4%	E: 3,483	5.3%	E: 3,679	5.6%	E: 3,423	5.4%	E: 17,632	5.4%
F: 2,629	4.0%	F: 2,512	4.0%	F: 2,589	3.9%	F: 2,511	3.8%	F: 2,389	3.8%	F: 12,630	3.9%
G: 5,614	8.6%	G: 5,214	8.2%	G: 5,512	8.3%	G: 5,553	8.5%	G: 5,136	8.1%	G: 27,029	8.3%
H: 802	1.2%	H: 820	1.3%	H: 802	1.2%	H: 868	1.3%	H: 767	1.2%	H: 4,059	1.3%
I: 3,721	5.7%	I: 3,575	5.6%	I: 3,610	5.5%	I: 3,500	5.3%	I: 3,656	5.7%	I: 18,062	5.6%
K: 3,655	5.6%	K: 3,799	6.0%	K: 4,002	6.0%	K: 3,718	5.7%	K: 3,517	5.5%	K: 18,691	5.8%
L: 5,781	8.9%	L: 5,572	8.8%	L: 5,978	9.0%	L: 5,824	8.9%	L: 5,747	9.0%	L: 28,902	8.9%
M: 1,511	2.3%	M: 1,372	2.2%	M: 1,510	2.3%	M: 1,413	2.2%	M: 1,420	2.2%	M: 7,226	2.2%
N: 2,357	3.6%	N: 2,465	3.9%	N: 2,526	3.8%	N: 2,696	4.1%	N: 2,492	3.9%	N: 12,536	3.9%
P: 2,890	4.4%	P: 2,901	4.6%	P: 2,987	4.5%	P: 2,856	4.4%	P: 2,889	4.5%	P: 14,523	4.5%
Q: 2,474	3.8%	Q: 2,499	3.9%	Q: 2,702	4.1%	Q: 2,583	3.9%	Q: 2,567	4.0%	Q: 12,825	4.0%
R: 2,744	4.2%	R: 2,626	4.1%	R: 2,751	4.2%	R: 2,802	4.3%	R: 2,888	4.5%	R: 13,811	4.3%
S: 4,012	6.1%	S: 3,818	6.0%	S: 4,081	6.2%	S: 4,175	6.4%	S: 3,955	6.2%	S: 20,041	6.2%
T: 3,831	5.9%	T: 3,763	5.9%	T: 3,921	5.9%	T: 3,876	5.9%	T: 3,851	6.0%	T: 19,242	5.9%
V: 4,875	7.5%	V: 4,710	7.4%	V: 4,743	7.2%	V: 4,830	7.4%	V: 4,703	7.4%	V: 23,861	7.4%
W: 665	1.0%	W: 697	1.1%	W: 675	1.0%	W: 704	1.1%	W: 765	1.2%	W: 3,506	1.1%
Y: 2,104	3.2%	Y: 2,125	3.3%	Y: 2,125	3.2%	Y: 2,225	3.4%	Y: 2,128	3.3%	Y: 10,707	3.3%
U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%
O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%
		X: 1	0.0%					X: 1	0.0%	X: 2	0.0%

Table G.2 Amino acid statistics for grouped T2SS proteins for k- fold cross validation

total			Group 1			Group 2			Group 3			Group 4			Group 5		
Length: 1,902			Length: 1,257			Length: 1,282			Length: 1,902			Length: 1,571			Length: 1,413		
Sequences: 668			Sequences: 134			Sequences: 134			Sequences: 134			Sequences: 133			Sequences: 133		
Lengths of 668 sequences:			Lengths of 134 sequences:			Lengths of 134 sequences:			Lengths of 134 sequences:			Lengths of 133 sequences:			Lengths of 133 sequences:		
Mean: 394.0	Std Dev. 219.4		Mean: 365.1	Std Dev. 206.8		Mean: 393.8	Std Dev. 183.6		Mean: 432.4	Std Dev. 246.5		Mean: 368.8	Std Dev. 228.6		Mean: 413.3	Std Dev. 217.4	
Minimum: 57	Maximum: 1902		Minimum: 57	Maximum: 1257		Minimum: 74	Maximum: 1282		Minimum: 92	Maximum: 1902		Minimum: 69	Maximum: 1571		Minimum: 96	Maximum: 1413	
Molecular weight (mean): 43.385 kDa			Molecular weight (mean): 40.308 kDa			Molecular weight (mean): 43.362 kDa			Molecular weight (mean): 47.781 kDa			Molecular weight (mean): 40.674 kDa			Molecular weight (mean): 45.219 kDa		
Isoelectric point (mean): 7.94			Isoelectric point (mean): 7.83			Isoelectric point (mean): 7.88			Isoelectric point (mean): 7.87			Isoelectric point (mean): 8.07			Isoelectric point (mean): 8.05		
Extinction Coefficient (mean): 36,927			Extinction Coefficient (mean): 34,389			Extinction Coefficient (mean): 34,351			Extinction Coefficient (mean): 41,338			Extinction Coefficient (mean): 39,118			Extinction Coefficient (mean): 35,460		
Freq	%		Freq	%	% non-ambig	Freq	%		Freq	%		Freq	%		Freq	%	
A: 24,886	9.5%		A: 4,599	9.4%	9.4%	A: 4,986	9.4%		A: 5,409	9.3%		A: 4,430	9.0%		A: 5,497	10.0%	
C: 1,792	0.7%		C: 372	0.8%	0.8%	C: 359	0.7%		C: 373	0.6%		C: 323	0.7%		C: 376	0.7%	
D: 13,584	5.2%		D: 2,499	5.1%	5.1%	D: 2,721	5.2%		D: 3,001	5.2%		D: 2,434	5.0%		D: 2,939	5.3%	
E: 15,678	6.0%		E: 2,914	6.0%	6.0%	E: 3,212	6.1%		E: 3,457	6.0%		E: 2,896	5.9%		E: 3,239	5.9%	
F: 9,505	3.6%		F: 1,802	3.7%	3.7%	F: 1,872	3.5%		F: 2,196	3.8%		F: 1,806	3.7%		F: 1,841	3.3%	
G: 18,695	7.1%		G: 3,438	7.0%	7.0%	G: 3,788	7.2%		G: 3,999	6.9%		G: 3,498	7.1%		G: 3,988	7.3%	
H: 3,879	1.5%		H: 805	1.6%	1.6%	H: 768	1.5%		H: 818	1.4%		H: 710	1.4%		H: 791	1.4%	
I: 16,593	6.3%		I: 3,139	6.4%	6.4%	I: 3,427	6.5%		I: 3,717	6.4%		I: 3,006	6.1%		I: 3,349	6.1%	
K: 11,248	4.3%		K: 2,003	4.1%	4.1%	K: 2,261	4.3%		K: 2,664	4.6%		K: 2,147	4.4%		K: 2,174	4.0%	
L: 29,972	11.4%		L: 5,599	11.4%	11.5%	L: 6,101	11.6%		L: 6,511	11.2%		L: 5,583	11.4%		L: 6,220	11.3%	
M: 6,869	2.6%		M: 1,236	2.5%	2.5%	M: 1,467	2.8%		M: 1,537	2.7%		M: 1,316	2.7%		M: 1,329	2.4%	
N: 9,398	3.6%		N: 1,756	3.6%	3.6%	N: 1,782	3.4%		N: 2,152	3.7%		N: 1,867	3.8%		N: 1,884	3.4%	
P: 11,626	4.4%		P: 2,146	4.4%	4.4%	P: 2,310	4.4%		P: 2,414	4.2%		P: 2,241	4.6%		P: 2,525	4.6%	
Q: 11,375	4.3%		Q: 2,025	4.1%	4.1%	Q: 2,094	4.0%		Q: 2,539	4.4%		Q: 2,336	4.8%		Q: 2,381	4.3%	
R: 16,824	6.4%		R: 3,182	6.5%	6.5%	R: 3,536	6.7%		R: 3,609	6.2%		R: 2,959	6.0%		R: 3,568	6.5%	
S: 17,946	6.8%		S: 3,395	6.9%	6.9%	S: 3,496	6.6%		S: 3,850	6.6%		S: 3,509	7.2%		S: 3,709	6.7%	
T: 14,463	5.5%		T: 2,620	5.4%	5.4%	T: 2,835	5.4%		T: 3,269	5.6%		T: 2,695	5.5%		T: 3,064	5.6%	
V: 19,738	7.5%		V: 3,540	7.2%	7.2%	V: 4,021	7.6%		V: 4,316	7.4%		V: 3,523	7.2%		V: 4,339	7.9%	
W: 2,743	1.0%		W: 494	1.0%	1.0%	W: 501	0.9%		W: 590	1.0%		W: 637	1.3%		W: 521	0.9%	
Y: 6,366	2.4%		Y: 1,256	2.6%	2.6%	Y: 1,227	2.3%		Y: 1,526	2.6%		Y: 1,129	2.3%		Y: 1,229	2.2%	
U: 2	0.0%		U: 14	0.0%	0.0%	U: 0	0.0%		U: 0	0.0%		U: 2	0.0%		U: 0	0.0%	
O: 0	0.0%		O: 33	0.1%	0.1%	O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%	
X: 1	0.0%		Z: 1	0.0%													
			X: 48	0.1%													
			B: 5	0.0%													

Table G.3 Amino acid statistics for grouped T3SS proteins for k- fold cross validation

TOTAL			Group 1			Group 2			Group 3			Group 4			Group 5		
Length: 2,674			Length: 795			Length: 914			Length: 2,674			Length: 767			Length: 1,282		
Sequences: 381			Sequences: 77			Sequences: 76			Sequences: 76			Sequences: 76			Sequences: 76		
Lengths of 381 sequences:			Lengths of 77 sequences:			Lengths of 76 sequences:			Lengths of 76 sequences:			Lengths of 76 sequences:			Lengths of 76 sequences:		
Mean: 303.4	Std Dev. 223.7		Mean: 288.1	Std Dev. 170.5		Mean: 304.0	Std Dev. 169.5		Mean: 346.1	Std Dev. 341.3		Mean: 257.3	Std Dev. 167.6		Mean: 321.7	Std Dev. 208.4	
Minimum: 60	Maximum: 2674		Minimum: 60	Maximum: 795		Minimum: 62	Maximum: 914		Minimum: 61	Maximum: 2674		Minimum: 67	Maximum: 767		Minimum: 62	Maximum: 1282	
Molecular weight (mean): 33.221 kDa			Molecular weight (mean): 31.561 kDa			Molecular weight (mean): 33.309 kDa			Molecular weight (mean): 37.674 kDa			Molecular weight (mean): 28.288 kDa			Molecular weight (mean): 35.292 kDa		
Isoelectric point (mean): 6.75			Isoelectric point (mean): 6.64			Isoelectric point (mean): 6.76			Isoelectric point (mean): 6.84			Isoelectric point (mean): 6.70			Isoelectric point (mean): 6.79		
Extinction Coefficient (mean): 27,250			Extinction Coefficient (mean): 24,591			Extinction Coefficient (mean): 27,873			Extinction Coefficient (mean): 30,310			Extinction Coefficient (mean): 24,790			Extinction Coefficient (mean): 28,722		
Freq	%		Freq	%		Freq	%		Freq	%		Freq	%		Freq	%	
A: 12,808	11.1%		A: 2,448	11.0%		A: 2,676	11.6%		A: 3,118	11.9%		A: 2,034	10.4%		A: 2,532	10.4%	
C: 986	0.9%		C: 200	0.9%		C: 224	1.0%		C: 223	0.8%		C: 184	0.9%		C: 155	0.6%	
D: 6,064	5.2%		D: 1,153	5.2%		D: 1,187	5.1%		D: 1,427	5.4%		D: 1,032	5.3%		D: 1,265	5.2%	
E: 6,686	5.8%		E: 1,225	5.5%		E: 1,417	6.1%		E: 1,448	5.5%		E: 1,128	5.8%		E: 1,468	6.0%	
F: 3,778	3.3%		F: 797	3.6%		F: 737	3.2%		F: 729	2.8%		F: 696	3.6%		F: 819	3.4%	
G: 7,510	6.5%		G: 1,406	6.3%		G: 1,532	6.6%		G: 1,739	6.6%		G: 1,289	6.6%		G: 1,544	6.3%	
H: 2,119	1.8%		H: 384	1.7%		H: 454	2.0%		H: 504	1.9%		H: 345	1.8%		H: 432	1.8%	
I: 6,484	5.6%		I: 1,351	6.1%		I: 1,175	5.1%		I: 1,363	5.2%		I: 1,178	6.0%		I: 1,417	5.8%	
K: 4,566	4.0%		K: 909	4.1%		K: 815	3.5%		K: 1,010	3.8%		K: 778	4.0%		K: 1,054	4.3%	
L: 13,462	11.6%		L: 2,615	11.8%		L: 2,699	11.7%		L: 3,027	11.5%		L: 2,259	11.6%		L: 2,862	11.7%	
M: 2,867	2.5%		M: 626	2.8%		M: 579	2.5%		M: 580	2.2%		M: 512	2.6%		M: 570	2.3%	
N: 4,258	3.7%		N: 864	3.9%		N: 779	3.4%		N: 941	3.6%		N: 723	3.7%		N: 951	3.9%	
P: 5,644	4.9%		P: 1,019	4.6%		P: 1,145	5.0%		P: 1,406	5.3%		P: 905	4.6%		P: 1,169	4.8%	
Q: 5,873	5.1%		Q: 1,078	4.9%		Q: 1,178	5.1%		Q: 1,408	5.4%		Q: 936	4.8%		Q: 1,273	5.2%	
R: 6,992	6.0%		R: 1,250	5.6%		R: 1,553	6.7%		R: 1,647	6.3%		R: 1,176	6.0%		R: 1,366	5.6%	
S: 7,946	6.9%		S: 1,564	7.1%		S: 1,505	6.5%		S: 1,715	6.5%		S: 1,394	7.1%		S: 1,768	7.2%	
T: 5,849	5.1%		T: 1,077	4.9%		T: 1,171	5.1%		T: 1,417	5.4%		T: 971	5.0%		T: 1,213	5.0%	
V: 8,036	7.0%		V: 1,514	6.8%		V: 1,561	6.8%		V: 1,819	6.9%		V: 1,341	6.9%		V: 1,801	7.4%	
W: 1,217	1.1%		W: 208	0.9%		W: 260	1.1%		W: 282	1.1%		W: 217	1.1%		W: 250	1.0%	
Y: 2,441	2.1%		Y: 496	2.2%		Y: 454	2.0%		Y: 497	1.9%		Y: 457	2.3%		Y: 537	2.2%	
U: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%	
O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%	

Table G.4 Amino acid statistics for grouped T4SS proteins for k- fold cross validation

TOTAL			GROUP 1			GROUP 2			GROUP 3			GROUP 4			GROUP 5		
Length: 3,095			Length: 1,934			Length: 1,891			Length: 2,179			Length: 3,095			Length: 2,282		
Sequences: 770			Sequences: 154			Sequences: 154			Sequences: 154			Sequences: 154			Sequences: 154		
Lengths of 770 sequences:			Lengths of 154 sequences:			Lengths of 154 sequences:			Lengths of 154 sequences:			Lengths of 154 sequences:			Lengths of 154 sequences:		
Mean: 445.9	Std Dev. 378.5		Mean: 457.3	Std Dev. 365.0		Mean: 443.9	Std Dev. 378.8		Mean: 433.1	Std Dev. 368.9		Mean: 447.6	Std Dev. 384.1		Mean: 447.5	Std Dev. 394.4	
Minimum: 47	Maximum: 3095		Minimum: 47	Maximum: 1934		Minimum: 56	Maximum: 1891		Minimum: 62	Maximum: 2179		Minimum: 65	Maximum: 3095		Minimum: 63	Maximum: 2282	
Molecular weight (mean): 49.166 kDa			Molecular weight (mean): 50.520 kDa			Molecular weight (mean): 48.977 kDa			Molecular weight (mean): 47.825 kDa			Molecular weight (mean): 49.273 kDa			Molecular weight (mean): 49.233 kDa		
Isoelectric point (mean): 7.36			Isoelectric point (mean): 7.42			Isoelectric point (mean): 7.38			Isoelectric point (mean): 7.48			Isoelectric point (mean): 7.36			Isoelectric point (mean): 7.14		
Extinction Coefficient (mean): 53,265			Extinction Coefficient (mean): 56,625			Extinction Coefficient (mean): 55,858			Extinction Coefficient (mean): 52,163			Extinction Coefficient (mean): 48,399			Extinction Coefficient (mean): 53,279		
Freq	%		Freq	%		Freq	%		Freq	%		Freq	%		Freq	%	
A: 31,274	9.1%		A: 6,448	9.2%		A: 6,401	9.4%		A: 5,880	8.8%		A: 6,311	9.2%		A: 6,234	9.0%	
C: 3,239	0.9%		C: 704	1.0%		C: 769	1.1%		C: 561	0.8%		C: 549	0.8%		C: 656	1.0%	
D: 20,041	5.8%		D: 4,064	5.8%		D: 4,046	5.9%		D: 3,962	5.9%		D: 3,917	5.7%		D: 4,052	5.9%	
E: 19,225	5.6%		E: 3,916	5.6%		E: 3,829	5.6%		E: 3,755	5.6%		E: 3,961	5.7%		E: 3,764	5.5%	
F: 12,270	3.6%		F: 2,703	3.8%		F: 2,296	3.4%		F: 2,318	3.5%		F: 2,500	3.6%		F: 2,453	3.6%	
G: 27,308	8.0%		G: 5,598	7.9%		G: 5,531	8.1%		G: 5,252	7.9%		G: 5,329	7.7%		G: 5,598	8.1%	
H: 7,085	2.1%		H: 1,633	2.3%		H: 1,472	2.2%		H: 1,309	2.0%		H: 1,310	1.9%		H: 1,361	2.0%	
I: 20,014	5.8%		I: 4,070	5.8%		I: 3,863	5.7%		I: 3,983	6.0%		I: 4,041	5.9%		I: 4,057	5.9%	
K: 16,611	4.8%		K: 3,321	4.7%		K: 3,189	4.7%		K: 3,284	4.9%		K: 3,465	5.0%		K: 3,352	4.9%	
L: 31,750	9.2%		L: 6,589	9.4%		L: 6,348	9.3%		L: 6,108	9.2%		L: 6,364	9.2%		L: 6,341	9.2%	
M: 7,189	2.1%		M: 1,424	2.0%		M: 1,382	2.0%		M: 1,492	2.2%		M: 1,449	2.1%		M: 1,442	2.1%	
N: 14,608	4.3%		N: 2,844	4.0%		N: 2,816	4.1%		N: 3,011	4.5%		N: 2,830	4.1%		N: 3,107	4.5%	
P: 15,079	4.4%		P: 3,160	4.5%		P: 3,053	4.5%		P: 2,811	4.2%		P: 3,058	4.4%		P: 2,997	4.3%	
Q: 14,718	4.3%		Q: 3,226	4.6%		Q: 2,867	4.2%		Q: 2,810	4.2%		Q: 2,977	4.3%		Q: 2,838	4.1%	
R: 19,807	5.8%		R: 4,011	5.7%		R: 4,173	6.1%		R: 3,842	5.8%		R: 3,917	5.7%		R: 3,864	5.6%	
S: 22,387	6.5%		S: 4,528	6.4%		S: 4,353	6.4%		S: 4,307	6.5%		S: 4,616	6.7%		S: 4,583	6.6%	
T: 22,005	6.4%		T: 4,252	6.0%		T: 4,187	6.1%		T: 4,450	6.7%		T: 4,540	6.6%		T: 4,576	6.6%	
V: 22,971	6.7%		V: 4,666	6.6%		V: 4,472	6.5%		V: 4,516	6.8%		V: 4,813	7.0%		V: 4,504	6.5%	
W: 4,328	1.3%		W: 950	1.3%		W: 902	1.3%		W: 862	1.3%		W: 744	1.1%		W: 870	1.3%	
Y: 11,428	3.3%		Y: 2,319	3.3%		Y: 2,414	3.5%		Y: 2,188	3.3%		Y: 2,236	3.2%		Y: 2,271	3.3%	
U: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%	
O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%	

Table G.5 Amino acid statistics for grouped T5SS proteins for k- fold cross validation

TOTAL			GROUP 1			GROUP 2			GROUP 3			GROUP 4			GROUP 5		
Length: 5,685			Length: 5,685			Length: 2,955			Length: 2,248			Length: 5,348			Length: 4,428		
Sequences: 221			Sequences: 45			Sequences: 44			Sequences: 44			Sequences: 44			Sequences: 44		
Lengths of 221 sequences:			Lengths of 45 sequences:			Lengths of 44 sequences:			Lengths of 44 sequences:			Lengths of 44 sequences:			Lengths of 44 sequences:		
Mean: 1246.0	Std Dev: 793.0		Mean: 1333.5	Std Dev. 987.3		Mean: 1095.5	Std Dev. 514.1		Mean: 1100.1	Std Dev. 387.9		Mean: 1421.2	Std Dev. 951.5		Mean: 1277.8	Std Dev. 870.5	
Minimum: 101	Maximum: 5685		Minimum: 101	Maximum: 5685		Minimum: 167	Maximum: 2955		Minimum: 526	Maximum: 2248		Minimum: 348	Maximum: 5348		Minimum: 148	Maximum: 4428	
Molecular weight (mean): 128.982 kDa			Molecular weight (mean): 137.303 kDa			Molecular weight (mean): 114.944 kDa			Molecular weight (mean): 114.695 kDa			Molecular weight (mean): 146.229 kDa			Molecular weight (mean): 131.551 kDa		
Isoelectric point (mean): 5.88			Isoelectric point (mean): 5.50			Isoelectric point (mean): 6.40			Isoelectric point (mean): 5.61			Isoelectric point (mean): 6.00			Isoelectric point (mean): 5.91		
Extinction Coefficient (mean): 114,263			Extinction Coefficient (mean): 115,691			Extinction Coefficient (mean): 109,655			Extinction Coefficient (mean): 109,009			Extinction Coefficient (mean): 118,301			Extinction Coefficient (mean): 118,625		
Freq	% A: 27,739	10.1%	Freq	%		Freq	%		Freq	%		Freq	%		Freq	%	
C: 469	0.2%		A: 6,125	10.2%		A: 4,820	10.0%		A: 4,606	9.5%		A: 6,056	9.7%		A: 6,132	10.9%	
D: 16,215	5.9%		C: 120	0.2%		C: 110	0.2%		C: 79	0.2%		C: 80	0.1%		C: 80	0.1%	
E: 9,239	3.4%		D: 3,273	5.5%		D: 2,907	6.0%		D: 2,763	5.7%		D: 3,785	6.1%		D: 3,487	6.2%	
F: 7,483	2.7%		E: 1,874	3.1%		E: 1,690	3.5%		E: 1,721	3.6%		E: 2,207	3.5%		E: 1,747	3.1%	
G: 34,887	12.7%		F: 1,662	2.8%		F: 1,331	2.8%		F: 1,441	3.0%		F: 1,686	2.7%		F: 1,363	2.4%	
H: 3,667	1.3%		G: 7,613	12.7%		G: 5,496	11.4%		G: 6,264	12.9%		G: 8,305	13.3%		G: 7,209	12.8%	
I: 14,162	5.1%		H: 723	1.2%		H: 801	1.7%		H: 704	1.5%		H: 765	1.2%		H: 674	1.2%	
K: 10,151	3.7%		I: 3,179	5.3%		I: 2,465	5.1%		I: 2,468	5.1%		I: 3,346	5.4%		I: 2,704	4.8%	
L: 20,667	7.5%		K: 1,873	3.1%		K: 1,894	3.9%		K: 1,865	3.9%		K: 2,268	3.6%		K: 2,251	4.0%	
M: 3,448	1.3%		L: 4,630	7.7%		L: 3,616	7.5%		L: 3,743	7.7%		L: 4,634	7.4%		L: 4,044	7.2%	
N: 20,682	7.5%		M: 756	1.3%		M: 605	1.3%		M: 632	1.3%		M: 849	1.4%		M: 606	1.1%	
P: 6,841	2.5%		N: 4,745	7.9%		N: 3,611	7.5%		N: 3,553	7.3%		N: 4,422	7.1%		N: 4,351	7.7%	
Q: 9,806	3.6%		P: 1,466	2.4%		P: 1,346	2.8%		P: 1,289	2.7%		P: 1,409	2.3%		P: 1,331	2.4%	
R: 8,465	3.1%		Q: 2,009	3.3%		Q: 1,826	3.8%		Q: 1,885	3.9%		Q: 2,189	3.5%		Q: 1,897	3.4%	
S: 25,231	9.2%		R: 1,699	2.8%		R: 1,739	3.6%		R: 1,549	3.2%		R: 1,744	2.8%		R: 1,734	3.1%	
T: 26,044	9.5%		S: 5,779	9.6%		S: 4,352	9.0%		S: 4,291	8.9%		S: 6,039	9.7%		S: 4,770	8.5%	
V: 20,016	7.3%		T: 6,047	10.1%		T: 4,287	8.9%		T: 4,226	8.7%		T: 5,912	9.5%		T: 5,572	9.9%	
W: 2,517	0.9%		V: 4,249	7.1%		V: 3,397	7.0%		V: 3,363	6.9%		V: 4,734	7.6%		V: 4,273	7.6%	
Y: 7,640	2.8%		W: 484	0.8%		W: 492	1.0%		W: 466	1.0%		W: 516	0.8%		W: 559	1.0%	
U: 0	0.0%		Y: 1,703	2.8%		Y: 1,418	2.9%		Y: 1,496	3.1%		Y: 1,586	2.5%		Y: 1,437	2.6%	
O: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%		U: 0	0.0%	
			O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%		O: 0	0.0%	

Table G.6 Amino acid statistics for grouped T6SS proteins for k- fold cross validation

TOTAL		GROUP 1		GROUP 2		GROUP 3		GROUP 4		GROUP 5	
Length: 1,335		Length: 1,206		Length: 1,335		Length: 1,166		Length: 1,137		Length: 1,193	
Sequences: 247		Sequences: 50		Sequences: 50		Sequences: 49		Sequences: 49		Sequences: 49	
Lengths of 247 sequences:		Lengths of 50 sequences:		Lengths of 50 sequences:		Lengths of 49 sequences:		Lengths of 49 sequences:		Lengths of 49 sequences:	
Mean: 448.6 Std Dev. 268.2		Mean: 477.1 Std Dev. 293.5		Mean: 477.9 Std Dev. 326.2		Mean: 404.8 Std Dev. 229.4		Mean: 424.3 Std Dev. 193.5		Mean: 457.9 Std Dev. 267.9	
Minimum: 90 Maximum: 1335		Minimum: 96 Maximum: 1206		Minimum: 137 Maximum: 1335		Minimum: 121 Maximum: 1166		Minimum: 154 Maximum: 1137		Minimum: 90 Maximum: 1193	
Molecular weight (mean): 50.149 kDa		Molecular weight (mean): 53.588 kDa		Molecular weight (mean): 53.609 kDa		Molecular weight (mean): 45.015 kDa		Molecular weight (mean): 47.602 kDa		Molecular weight (mean): 50.792 kDa	
Isoelectric point (mean): 6.47		Isoelectric point (mean): 6.55		Isoelectric point (mean): 6.48		Isoelectric point (mean): 6.26		Isoelectric point (mean): 6.32		Isoelectric point (mean): 6.74	
Extinction Coefficient (mean): 55,724		Extinction Coefficient (mean): 64,801		Extinction Coefficient (mean): 58,136		Extinction Coefficient (mean): 45,842		Extinction Coefficient (mean): 52,144		Extinction Coefficient (mean): 57,464	
Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
A: 9,612	8.7%	A: 1,939	8.1%	A: 1,978	8.3%	A: 1,777	9.0%	A: 1,755	8.4%	A: 2,163	9.6%
C: 1,029	0.9%	C: 206	0.9%	C: 213	0.9%	C: 188	0.9%	C: 230	1.1%	C: 192	0.9%
D: 6,395	5.8%	D: 1,381	5.8%	D: 1,391	5.8%	D: 1,122	5.7%	D: 1,210	5.8%	D: 1,291	5.8%
E: 7,097	6.4%	E: 1,511	6.3%	E: 1,532	6.4%	E: 1,286	6.5%	E: 1,339	6.4%	E: 1,429	6.4%
F: 4,497	4.1%	F: 988	4.1%	F: 966	4.0%	F: 792	4.0%	F: 921	4.4%	F: 830	3.7%
G: 7,305	6.6%	G: 1,542	6.5%	G: 1,492	6.2%	G: 1,348	6.8%	G: 1,284	6.2%	G: 1,639	7.3%
H: 2,584	2.3%	H: 542	2.3%	H: 532	2.2%	H: 495	2.5%	H: 508	2.4%	H: 507	2.3%
I: 5,126	4.6%	I: 1,168	4.9%	I: 1,163	4.9%	I: 911	4.6%	I: 876	4.2%	I: 1,008	4.5%
K: 4,388	4.0%	K: 980	4.1%	K: 1,074	4.5%	K: 718	3.6%	K: 765	3.7%	K: 851	3.8%
L: 12,653	11.4%	L: 2,660	11.2%	L: 2,663	11.1%	L: 2,366	11.9%	L: 2,446	11.8%	L: 2,518	11.2%
M: 2,335	2.1%	M: 499	2.1%	M: 557	2.3%	M: 373	1.9%	M: 435	2.1%	M: 471	2.1%
N: 3,963	3.6%	N: 954	4.0%	N: 909	3.8%	N: 675	3.4%	N: 732	3.5%	N: 693	3.1%
P: 5,850	5.3%	P: 1,200	5.0%	P: 1,196	5.0%	P: 1,084	5.5%	P: 1,183	5.7%	P: 1,187	5.3%
Q: 5,646	5.1%	Q: 1,216	5.1%	Q: 1,211	5.1%	Q: 1,052	5.3%	Q: 1,080	5.2%	Q: 1,087	4.8%
R: 7,224	6.5%	R: 1,524	6.4%	R: 1,516	6.3%	R: 1,265	6.4%	R: 1,372	6.6%	R: 1,547	6.9%
S: 7,522	6.8%	S: 1,658	7.0%	S: 1,624	6.8%	S: 1,401	7.1%	S: 1,393	6.7%	S: 1,446	6.4%
T: 5,762	5.2%	T: 1,264	5.3%	T: 1,264	5.3%	T: 970	4.9%	T: 1,047	5.0%	T: 1,217	5.4%
V: 7,004	6.3%	V: 1,474	6.2%	V: 1,562	6.5%	V: 1,222	6.2%	V: 1,325	6.4%	V: 1,421	6.3%
W: 1,629	1.5%	W: 379	1.6%	W: 331	1.4%	W: 264	1.3%	W: 304	1.5%	W: 351	1.6%
Y: 3,186	2.9%	Y: 768	3.2%	Y: 721	3.0%	Y: 526	2.7%	Y: 584	2.8%	Y: 587	2.6%
U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%
O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%

Table G.7 Amino acid statistics for grouped orphan-secreted proteins for k-fold cross validation

TOTAL		GROUP 1		GROUP 2		GROUP 3		GROUP 4		GROUP 5	
Length: 5,468		Length: 3,577		Length: 3,321		Length: 5,468		Length: 4,342		Length: 5,020	
Sequences: 2,533		Sequences: 507		Sequences: 507		Sequences: 507		Sequences: 506		Sequences: 506	
Lengths of 2,533 sequences:		Lengths of 507 sequences:		Lengths of 507 sequences:		Lengths of 507 sequences:		Lengths of 506 sequences:		Lengths of 506 sequences:	
Mean: 383.8 Std Dev. 370.3		Mean: 409.9 Std Dev. 376.3		Mean: 378.9 Std Dev. 315.6		Mean: 403.3 Std Dev. 457.6		Mean: 355.2 Std Dev. 318.4		Mean: 371.8 Std Dev. 362.8	
Minimum: 26 Maximum: 5468		Minimum: 26 Maximum: 3577		Minimum: 40 Maximum: 3321		Minimum: 32 Maximum: 5468		Minimum: 39 Maximum: 4342		Minimum: 38 Maximum: 5020	
		Molecular weight (mean): 44.135 kDa		Molecular weight (mean): 40.745 kDa		Molecular weight (mean): 43.272 kDa		Molecular weight (mean): 38.194 kDa		Molecular weight (mean): 40.112 kDa	
Isoelectric point (mean): 6.79		Isoelectric point (mean): 6.87		Isoelectric point (mean): 6.79		Isoelectric point (mean): 6.80		Isoelectric point (mean): 6.79		Isoelectric point (mean): 6.71	
		Extinction Coefficient (mean): 48,661		Extinction Coefficient (mean): 47,089		Extinction Coefficient (mean): 46,645		Extinction Coefficient (mean): 42,090		Extinction Coefficient (mean): 45,753	
Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
A: 104,757	10.8%	A: 21,616	10.4%	A: 20,840	10.8%	A: 21,844	10.7%	A: 20,148	11.2%	A: 20,309	10.8%
C: 7,221	0.7%	C: 1,477	0.7%	C: 1,396	0.7%	C: 1,472	0.7%	C: 1,365	0.8%	C: 1,511	0.8%
D: 62,187	6.4%	D: 13,463	6.5%	D: 12,081	6.3%	D: 13,230	6.5%	D: 11,496	6.4%	D: 11,917	6.3%
E: 48,058	4.9%	E: 10,392	5.0%	E: 9,337	4.9%	E: 10,123	5.0%	E: 8,982	5.0%	E: 9,224	4.9%
F: 31,160	3.2%	F: 6,764	3.3%	F: 6,030	3.1%	F: 6,427	3.1%	F: 5,577	3.1%	F: 6,362	3.4%
G: 86,188	8.9%	G: 18,336	8.8%	G: 17,066	8.9%	G: 18,372	9.0%	G: 16,011	8.9%	G: 16,403	8.7%
H: 16,814	1.7%	H: 3,488	1.7%	H: 3,390	1.8%	H: 3,227	1.6%	H: 3,381	1.9%	H: 3,328	1.8%
I: 45,446	4.7%	I: 9,964	4.8%	I: 8,686	4.5%	I: 9,847	4.8%	I: 8,099	4.5%	I: 8,850	4.7%
K: 42,024	4.3%	K: 9,225	4.4%	K: 8,295	4.3%	K: 8,693	4.3%	K: 7,469	4.2%	K: 8,342	4.4%
L: 80,226	8.3%	L: 16,958	8.2%	L: 15,589	8.1%	L: 16,623	8.1%	L: 15,315	8.5%	L: 15,741	8.4%
M: 16,590	1.7%	M: 3,420	1.6%	M: 3,294	1.7%	M: 3,472	1.7%	M: 3,106	1.7%	M: 3,298	1.8%
N: 44,865	4.6%	N: 10,233	4.9%	N: 8,558	4.5%	N: 9,810	4.8%	N: 7,654	4.3%	N: 8,610	4.6%
P: 48,974	5.0%	P: 10,376	5.0%	P: 9,802	5.1%	P: 10,393	5.1%	P: 9,101	5.1%	P: 9,302	4.9%
Q: 36,926	3.8%	Q: 7,866	3.8%	Q: 7,196	3.7%	Q: 7,794	3.8%	Q: 6,948	3.9%	Q: 7,122	3.8%
R: 46,285	4.8%	R: 9,377	4.5%	R: 9,411	4.9%	R: 9,506	4.6%	R: 9,057	5.0%	R: 8,934	4.7%
S: 70,769	7.3%	S: 15,101	7.3%	S: 14,328	7.5%	S: 14,873	7.3%	S: 12,825	7.1%	S: 13,642	7.3%
T: 72,010	7.4%	T: 15,517	7.5%	T: 14,355	7.5%	T: 15,475	7.6%	T: 12,938	7.2%	T: 13,725	7.3%
V: 68,881	7.1%	V: 15,000	7.2%	V: 13,881	7.2%	V: 14,639	7.2%	V: 12,453	6.9%	V: 12,908	6.9%
W: 13,062	1.3%	W: 2,697	1.3%	W: 2,746	1.4%	W: 2,669	1.3%	W: 2,395	1.3%	W: 2,555	1.4%
Y: 29,808	3.1%	Y: 6,549	3.2%	Y: 5,836	3.0%	Y: 5,967	2.9%	Y: 5,404	3.0%	Y: 6,052	3.2%
U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%
O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%
X: 4	0.0%	X: 3	0.0%			X: 1	0.0%			B: 1	0.0%
B: 1	0.0%										

Table G.8 Amino acid statistics for grouped LPTXG motifed proteins for k- fold cross validation

Group 1		Group 2		Group 3		Group 4		Group 5		total	
Length: 1,913		Length: 2,987		Length: 2,963		Length: 2,243		Length: 2,972		Length: 2,987	
Sequences: 56		Sequences: 56		Sequences: 55		Sequences: 55		Sequences: 55		Sequences: 277	
Lengths of 56 sequences:		Lengths of 56 sequences:		Lengths of 55 sequences:		Lengths of 55 sequences:		Lengths of 55 sequences:		Lengths of 277 sequences:	
Mean: 585.3	Std Dev. 414.9	Mean: 696.5	Std Dev. 557.7	Mean: 671.7	Std Dev. 465.9	Mean: 662.5	Std Dev. 532.8	Mean: 673.0	Std Dev. 537.3	Mean: 657.7	Std Dev. 505.9
Minimum: 78	Maximum: 1913	Minimum: 49	Maximum: 2987	Minimum: 50	Maximum: 2963	Minimum: 44	Maximum: 2243	Minimum: 48	Maximum: 2972	Minimum: 44	Maximum: 2987
Molecular weight (mean): 63.335 kDa		Molecular weight (mean): 74.967 kDa		Molecular weight (mean): 71.523 kDa		Molecular weight (mean): 71.791 kDa		Molecular weight (mean): 72.521 kDa		Molecular weight (mean): 70.815 kDa	
Isoelectric point (mean): 6.37		Isoelectric point (mean): 6.69		Isoelectric point (mean): 6.67		Isoelectric point (mean): 6.65		Isoelectric point (mean): 7.00		Isoelectric point (mean): 6.67	
Extinction Coefficient (mean): 58,438		Extinction Coefficient (mean): 71,158		Extinction Coefficient (mean): 62,355		Extinction Coefficient (mean): 63,889		Extinction Coefficient (mean): 62,713		Extinction Coefficient (mean): 63,719	
Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
A: 3,027	9.2%	A: 3,629	9.3%	A: 3,577	9.7%	A: 3,065	8.4%	A: 3,074	8.3%	A: 16,372	9.0%
C: 117	0.4%	C: 213	0.5%	C: 211	0.6%	C: 153	0.4%	C: 89	0.2%	C: 783	0.4%
D: 2,036	6.2%	D: 2,620	6.7%	D: 2,537	6.9%	D: 2,485	6.8%	D: 2,434	6.6%	D: 12,112	6.6%
E: 1,987	6.1%	E: 2,370	6.1%	E: 1,882	5.1%	E: 2,416	6.6%	E: 2,162	5.8%	E: 10,817	5.9%
F: 988	3.0%	F: 1,168	3.0%	F: 1,050	2.8%	F: 1,070	2.9%	F: 1,117	3.0%	F: 5,393	3.0%
G: 2,379	7.3%	G: 3,145	8.1%	G: 3,157	8.5%	G: 2,731	7.5%	G: 3,093	8.4%	G: 14,505	8.0%
H: 353	1.1%	H: 437	1.1%	H: 467	1.3%	H: 423	1.2%	H: 420	1.1%	H: 2,100	1.2%
I: 1,710	5.2%	I: 1,834	4.7%	I: 1,682	4.6%	I: 1,798	4.9%	I: 1,894	5.1%	I: 8,918	4.9%
K: 2,679	8.2%	K: 3,084	7.9%	K: 2,776	7.5%	K: 3,072	8.4%	K: 3,207	8.7%	K: 14,818	8.1%
L: 2,219	6.8%	L: 2,493	6.4%	L: 2,261	6.1%	L: 2,325	6.4%	L: 2,370	6.4%	L: 11,668	6.4%
M: 486	1.5%	M: 491	1.3%	M: 467	1.3%	M: 483	1.3%	M: 453	1.2%	M: 2,380	1.3%
N: 2,065	6.3%	N: 2,288	5.9%	N: 2,123	5.7%	N: 2,089	5.7%	N: 2,373	6.4%	N: 10,938	6.0%
P: 1,694	5.2%	P: 1,742	4.5%	P: 1,851	5.0%	P: 1,795	4.9%	P: 1,832	4.9%	P: 8,914	4.9%
Q: 1,083	3.3%	Q: 1,272	3.3%	Q: 1,251	3.4%	Q: 1,314	3.6%	Q: 1,212	3.3%	Q: 6,132	3.4%
R: 809	2.5%	R: 898	2.3%	R: 856	2.3%	R: 796	2.2%	R: 819	2.2%	R: 4,178	2.3%
S: 2,242	6.8%	S: 2,843	7.3%	S: 2,650	7.2%	S: 2,515	6.9%	S: 2,403	6.5%	S: 12,653	6.9%
T: 3,225	9.8%	T: 3,828	9.8%	T: 3,973	10.8%	T: 3,568	9.8%	T: 3,735	10.1%	T: 18,329	10.1%
V: 2,306	7.0%	V: 2,895	7.4%	V: 2,760	7.5%	V: 2,800	7.7%	V: 2,851	7.7%	V: 13,612	7.5%
W: 304	0.9%	W: 339	0.9%	W: 328	0.9%	W: 302	0.8%	W: 311	0.8%	W: 1,584	0.9%
Y: 1,070	3.3%	Y: 1,415	3.6%	Y: 1,083	2.9%	Y: 1,238	3.4%	Y: 1,164	3.1%	Y: 5,970	3.3%
U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%	U: 0	0.0%
O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%	O: 0	0.0%

APPENDIX H

SECRETION RELATED PATTERNS

Table H.1 TISS-1-15 patterns. A. Highly abundant pattern hits in TISS-1-15. B. The most abundant pattern and its pairs in TISS-1-15.

A

Patterns	Counts
EINAAGGIDGRKVEL	17
MAVSEINAKGGVLGK	16
GAELAVKEINEAGGI	16
KIDVVIAGMTATEER	15
EINAAGGIHGRQLEI	15
QINARGGILGRPIEL	15
INDAGGINGRPLEII	15
TEERKKKYDFSEPYF	15
DIDLAKAIAKELGVK	14
GAELAVQRINEAGGI	14
ITEERKKKYDFSEPY	14
QINAGGGIGGRPLEL	14
VEDLKGKKVGVQTGS	14
AVNAAGGINGRKIEL	14
VNAAGGINGRKIELK	14
LALDEINAAGGIDGR	14
INAAGGIDGRKVELK	14
QINGAGGVLGRPLEL	14
KMAVSEINAKGGVLG	13
TPERKKQVDFSDSYM	13
DGLIPALESKKFDVI	13
KVDVVIAGMTETPER	13
EMAVAQINAAGGVLG	13
MTITDERKQQMDFSD	13
INAAGGIHGRQLEII	13
TEERKQSVDFSEVYY	13
INARGGILGRPIELL	13

B

EINAAGGIDGRKVEL pattern pairs
EINAKGGVLGKKLEV
EINAAGGIHGRQLEI
QINARGGILGRPIEL
QINAAGGHKGRPVEL
EINAKGGLLGKQVEL
RINDAGGINGRPLEI
DINAAGGVLDQPVEL
EINEAGGINGAQIEF
QINAGGGIGGRPLEL
AVNRAGGVNGRKIEL
AVNAAGGINGRKIEL
EINAAGGINGKKLEV
EINAAGGIDGRKVEL
DINGEGGIGGRKITL
ELNDAGGIDGRPVEL
AINAAGGVNGKKLEM
QINGAGGVLGRPLEL

Table H.2 TISS-2-15 patterns. A. Highly abundant pattern hits in TISS-2-15. B. The most abundant pattern and its pairs in TISS-2-15.

A

Patterns	Counts
KLAIEEINAAGGVLG	21
LAIEEINAAGGVLGK	20
LAVEEINASGGILGR	19
AIEEINAAGGVLGKQ	19
GLELAAEEINAAGGI	19
LAAEEINAAGGILGR	19
ELAVEEINASGGILG	18
EIAAEEINAAGGVNG	18
IAAEEINAAGGVNGK	18
ELAAEEINAAGGILG	18
GVELAVEEINASGGI	17
NGLELAAEEINAAGG	17
AAEEINAAGGILGRP	17
AEEINAAGGILGRPV	17
AVEEINASGGILGRK	16
LAIDEINAAGGVLGR	16
DEINAAGGVLGREIE	16
TPERKKEVDFSVPYY	16
QYAIDEINAAGGIAG	16
GAEIAAEEINAAGGV	16
AAEEINAAGGVNGKR	16
LELAAEEINAAGGIL	16
EEINAAGGILGRPVE	16
VELAVEEINASGGIL	15
ITYERKKVVDFSIPY	15
YAIDEINAAGGIAGS	15
INAAGGINGKKLETV	15
EEINAAGGVLGKQIE	15
AEIAAEEINAAGGVN	15
EINAAGGILGRPVEI	15

B

KLAIEEINAAGGVLG pattern pairs
ELAVEEINASGGILG
KAAVAEANAAGGING
KQAGAEALMAGGVLG
LLAIDEINAAGGVLG
RLIVEQYNARGGVLG
QYAIDEINAAGGIAG
EMAVEEINSAGGVVL
DTAVERINAAGGING
KLAIEQINASGGVLG
KLAIEEINAAGGVLG
LAAAEAINAAGGAAG
TIALEKINASGGVMG
ELAAQEINAQGGVFV
EMAVEEINAAGGIKS
EIAAEEINAAGGVNG
NLGIKEVNDAGGVLG
DLALDEIQAAGGVNG
ELAIDEINKAGGIDG
QLAVNEINDKGGVLG
LIRIDEINAAGGLLG
ELAAEEINAAGGILG

Table H.3 TISS-3-15 patterns. A. Highly abundant pattern hits in TISS-3-15. B. The most abundant pattern and its pairs in TISS-3-15.

A

Patterns	Counts
LAIEEINAKGGVLGR	23
DLAIEEINAKGGVLG	21
AIEEINAKGGVLGRK	21
IMAIEEINAAGGVNG	19
MAIEEINAAGGVNGM	19
LAVEEVNAAGGLLGR	18
GIDLAIEEINAKGGV	18
IEEINAKGGVLGRKL	18
EEINAKGGVLGRKLE	18
LAIDEINAQGGLLGR	17
QLAADEINAAGGING	17
AIEEINAAGGVNGME	17
LAASEINAAGGVLGQ	17
EINAAGGIDGRKVEL	17
LKADGTLAKLSQKWF	16
DGTLAKLSQKWFGED	16
LLAIDEINAQGGLLG	16
GFQLAADEINAAGGI	16
LAADEINAAGGINGS	16
KLAASEINAAGGVLG	16
AASEINAAGGVLGQQ	16
EINAAGGVLGQQIEL	16
MAVSEINAKGGVLGK	16
GAELAVKEINEAGGI	16
AIDEINAQGGLLGRR	15
LAVDEINKAGGVMGR	15
EINAKGGVLGRKLEV	15
TEEREKVIDFTAPYY	15
AKLAASEINAAGGVL	15
GVQMAIDEANAAGGV	15
KIDVVIAGMTATEER	15
EINAAGGIHGRQLEI	15
QINARGGILGRPIEL	15
INDAGGINGRPLEII	15
TEERKKKYDFSEPYF	15

B

LAIEEINAKGGVLGR pattern pairs
LAIAEINARGGIGGR
LAVQEINKAGGVDGK
MRFEEANAAGGVHGR
LAVEEVNAAGGLLGR
MAIADLNAAGGVLGK
IAVDEINAAGGVQGE
LAIDEINAQGGLLGR
FAIEEINADGGILDQ
LAVDEINKAGGVMGR
LALEEQNARGGINGR
LAVDQINASGGLLGR
LAADEINAAGGINGS
MVVERINAEGGVLGR
LYIEELNAQGGIHGV
LAVDQINANGGVLDR
MAIEEINAAGGVNGM
QAIADINAKGGIKGD
AAFEEANRAGGVLGR
MAIDDINAKGGVLGK
LAIEEINAKGGVLGR
YAISLINKDGGVLGK
LAASEINAAGGVLGQ
MAIDEANAAGGVAGD

Table H.4 TISS-4-15 patterns. A. Highly abundant pattern hits in TISS-4-15.

B. The most abundant pattern and its pairs in TISS-4-15

A

Patterns	Count
LAIDEINAAGGLLGR	22
ITDERKKKYDFSDPY	20
QLAIDEINAAGGLLG	19
LAVEQINAAGGVLGR	19
GKLVGFDVDIARAVA	18
LAVKEINDAGGIDGR	18
LAVDEVNAAGGVHGR	17
AIDEINAAGGLLGRS	16
ELAVEQINAAGGVLG	16
LAVSEINRCGGILGR	16
SEINAAGGVLGRPVE	16
EINAAGGVLGRPVEV	16
ITEERKKVVDFTVPY	16
GKLTGFEVDLAKAVA	16
LATEEINNAGGVLGR	16
EINNAGGVLGRKIEL	16
GAELAVKEINDAGGI	15
TDERKKKYDFSDPYI	15
MAVSEINAAGGVLGR	15
AVSEINAAGGVLGRP	15
VSEINAAGGVLGRPV	15
TEERKKVVDFTVPYY	15
ENGKLTGFEVDLAKA	15
EEINNAGGVLGRKIE	15
INNAGGVLGRKIELL	15

B

LAIDEINAAGGLLGR pattern pairs
MTIDQINAKGGVLGR
LAIDEINAAGGLLGR
LAVKEINDAGGIDGR
LALDEINTVGGIHGR
LAVEQINAAGGVLGR
LAVKDINAAGGVLGK
LAIEEINAAGGLKLN
LLIENQNAAGGLLGC
QAVSDINQAGGILGR
LAVSEINRCGGILGR
LLVKEINGAGGVDGR
LAIEEINNDGGIDGK
MAVSEINAAGGVLGR
AWADYVNARGGLLGR
LAVHDINNTGGVLGR
LMAEEWNAKGGVLGR
LAVDEVNAAGGVHGR
IAAAEINAAGGIHGR
LAIKEINENGGIFGK
LATEEINNAGGVLGR
LAADEVNKAGGVLGK
LAAAELNAANGIGGR

Table H.5 TISS-5-15 patterns. A. Highly abundant pattern hits in TISS-5-15. B. The most abundant pattern and its pairs in TISS-5-15

A

Patterns	Counts
EINAAGGIDGRKVEL	17
MAVSEINAKGGVLGK	16
GAELAVKEINEAGGI	16
KIDVVIAGMTATEER	15
EINAAGGIHGRQLEI	15
QINARGGILGRPIEL	15
INDAGGINGRPLEII	15
TEERKKKYDFSEPYF	15
DIDLAKAIAKELGVK	14
GAELAVQRINEAGGI	14
ITEERKKKYDFSEPY	14
QINAGGGIGGRPLEL	14
VEDLKGKKVGVQTGS	14
AVNAAGGINGRKIEL	14
VNAAGGINGRKIELK	14
LALDEINAAGGIDGR	14
INAAGGIDGRKVELK	14
QINGAGGVLGRPLEL	14
KMAVSEINAKGGVLG	13
TPERKKQVDFSDSYM	13
DGLIPALESKKFDVI	13
KVDVVIAGMTETPER	13
EMAVAQINAAGGVLG	13
MTITDERKQQMDFSD	13
INAAGGIHGRQLEII	13
TEERKQSVDFSEVYY	13
INARGGILGRPIELL	13
RINDAGGINGRPLEI	13
INAGGGIGGRPLELI	13
GFDLALDEINAAGGI	13
INGAGGVLGRPLELY	13

B

EINAAGGIDGRKVEL pattern pairs
EINAKGGVLGKKLEV
EINAAGGIHGRQLEI
QINARGGILGRPIEL
QINAAGGHKGRPVEL
EINAKGGLLGKQVEL
RINDAGGINGRPLEI
DINAAGGVLDQPVEL
EINEAGGINGAQIEF
QINAGGGIGGRPLEL
AVNRAGGVNGRKIEL
AVNAAGGINGRKIEL
EINAAGGINGKKLEV
EINAAGGIDGRKVEL
DINGEGGIGGRKITL
ELNDAGGIDGRPVEL
AINAAGGVNGKKLEM
QINGAGGVLGRPLEL

Table H.6 TISS-total-15 patterns. A. Highly abundant pattern hits in whole TISS. B. The most abundant pattern and its pairs in whole TISS data

A

Patterns	Counts
LAIDEINAAGGVLGR	96
LAVEEINAKGGILGR	94
LAAEEINAAGGILGR	94
LAIEEINAAGGVDGR	91
KLAIEEINAAGGVLG	90
LAVEEINASGGILGR	89
ELAAEEINAAGGILG	89
LAIDEINAAGGLLGR	88
LAVEQINAAGGVLGR	88
LAVDEVNAAGGVHGR	88
ELALDEINAAGGVLG	88
LAIEEINAAGGVLGK	87
LAIEEINAKGGVLGR	87
ELAVEEINASGGILG	86
ELAVEEINAKGGILG	86
ITDERKKVFDFSDPY	84
ITDERKKKYDFSDPY	84
ELAVEQINAAGGVLG	84
LAVEEVNAAGGLLGR	84
DEINAAGGVLGREIE	81
GFDIDIANAIAKKLG	81
LALDEINAAGGVLGG	81
LLAIDEINAAGGVLG	80
MAVSEINAAGGVLGR	80
LALDEINAAGGIDGR	80
AAEEINAAGGILGRP	80

B

LAIDEINAAGGVLGR pairs of first 25
LAVEEINASGGILGR
MAVSEINAKGGVLGK
LAVDHINQDGGLLGR
AAVAEANAAGGINGR
LWAKDFNAAGGVCGR
LAIDEINAAGGVLGR
LAIAEINARGGIGGR
MTIDQINAKGGVLGR
LAVAKINAQGGVLGQ
LAIDEINAAGGLLGR
LAVKEINDAGGIDGR
MAVAQINAAGGVLGE
LALDEINTVGGIHGR
LAIDQANAAGGIQGR
LAVEQINAAGGVLGR
LAVQEINKAGGVDGK
FAIQQINAAGGVDGR
LAVKDINAAGGVLGK
LAIEEINAAGGVDGR
AHFDEINANGGIHGR
MRFEEANAAGGVHGR
MRVDEINAAGGIHGR
LAIEEINAAGGLKLN
LLIENQNAAGGLLGC
YAIDEINAAGGIAGS
IAEADINAAGGVLGC

H.2 Type 2 Secretion Related Patterns

Table H.7 T2SS-1-15 patterns. A. Highly abundant pattern hits in T2SS-1-15. B. The most abundant pattern and its pairs in T2SS-1-15.

A

Patterns	Counts
ILVSGPTGSGKTTLL	17
IVICGPTGSGKTTTL	17
VLLTGPTGSGKTTTL	16
LTGPTGSGKTTTLYA	16
VSGPTGSGKTTLLNA	16
MLISGGTGSGKTTLL	16
LISGGTGSGKTTLLN	16
ILVTGPTGSGKTTTL	16
LVTGPTGSGKTTTLY	16
VTGPTGSGKTTTLYA	16
GPTGSGKTTTLYAMI	16
ILVSGPTGAGKTTLL	16
LVSGPTGAGKTTLLL	16
VSGGTGAGKTTLLNA	16
LLVTGPTGSGKTTTL	16
LVTGPTGSGKTTTLY	16
LLTGPTGSGKTTTLY	15
TGPTGSGKTTTLYAM	15
GPTGSGKTTTLYAML	15
LVSGPTGSGKTTLLN	15
SGPTGSGKTTLLNAL	15
VICGPTGSGKTTTLY	15
ICGPTGSGKTTTLYA	15
GPTGSGKTTTLYAAL	15
VLFAGPTGVGKTTLL	15
LFAGPTGVGKTTLLN	15
TGPTGSGKTTTLYAM	15
IILVSGPTGAGKTTL	15
LLVTGPTGSGKSTTL	15
LVTGPTGSGKSTTLY	15
VTGPTGSGKSTTLYA	15

B

ILVSGPTGSGKTTLL	Pattern Pairs
IVISGAPGSGKSTLA	
VLLTGPTGSGKTTTL	
ILVSGPTGSGKTTLL	
YLFSGPVGSGKTTTM	
LLIGGATGAGKTTVL	
IVICGPTGSGKTTTL	
ILVAGDVGAGKTSLL	
MLISGGTGSGKTTLL	
VLFAGPTGVGKTTLL	
ILVTGPTGSGKTTTL	
ILVSGPTGAGKTTLL	
MIVAGGTASGKTTTL	
ILMTGPTGSGKTVSL	
VVVSGGTGAGKTTLL	
LLVTGPTGSGKSTTL	
IILTGPTGSGKSKSL	
LLVTGPTGSGKTTTL	

Table H.8 T2SS-2-15 patterns. A. Highly abundant pattern hits in T2SS-2-15. B. The most abundant pattern and its pairs in T2SS-2-15.

A

Patterns	Counts
VLVSGPTGSGKTTTL	30
LVSGPTGSGKTTTLY	29
NILVSGGTGSGKTTL	29
ILVSGGTGSGKTTLM	29
VLVTGPTGSGKTTTL	29
VLVTGPTGSGKTTTL	29
LLVTGPTGSGKTTTL	28
LVTGPTGSGKTTTLY	28
IVLVSGPTGSGKTTT	28
LVTGPTGSGKTTTLY	28
LVTGPTGSGKTTTLY	28
LNILVSGGTGSGKTT	27
LMVAGGTGSGKTTSL	27
IVLVTGPTGSGKTTT	27
ILVTGPTGSGKSTTL	26
GIVLVSGPTGSGKTT	26
LRQDPDVIMVGEIRD	26
LVSGGTGSGKTTLMN	26
MVAGGTGSGKTTSLN	26
LRQRPDIIIVGETRG	26
MVLVTGPTGSGKTTT	26
SLRYRPDMIVVGEIR	25
LRYRPDMIVVGEIRG	25
ILLVTGPTGSGKTTT	25
IISGGTGSGKTTTLN	25
LLVTGATGSGKSTTL	25
LRQDPDIIMVGEIRD	25
GGTGSGKTTTLNTIA	25
FLVTGGTGAGKTTLL	25
LVTGGTGAGKTTLLS	25

B

VLVSGPTGSGKTTTL pattern pairs
VIISGGTGSGKTTLL
LLFSGPTGSGKSTLM
LLVTGPTGSGKTTTL
FIISGGTGSGKTTTL
ILISGETGTGKTEVL
ILVTGPTGSGKSTTL
LLVTGATGSGKSTTL
IIVTGPTGSGKSTTL
ILVTGPTGSGKTVSL
YLIVGSTGSGKTTFL
VLVSGPTGSGKTTTL
AIIAGGTGSGKTTTL
FIMAGQTGSGKTTTI
VLAIGPAGSGRTASL
FLVTGGTGAGKTTLL
YLFAGPVGSGKTTLM
FLVTGGTASGKTSML
VLVTGPTGSGKSTTL
ILVSGGTGSGKTTLM
ILLCGPTGSGKSTTV
LMVAGGTGSGKTTSL
IVLSGPTGSGKSTTL
VLVVGSTGSGKSTSL
MLVAGGTASGKTTAL
VFVVGETASGKTTTL
ILVSGGTSSGKTSLL
VLVTGPTGSGKTTTL
VLITGGTGAGKTTLL
VLVTGPTGSGKTTTL
ILISGGTGTGKTTLL

Table H.9 T2SS-3-15 patterns. A. Highly abundant pattern hits in T2SS-3-15. B. The most abundant pattern and its pairs in T2SS-3-15.

A

Patterns	Counts
VTGPTGSGKTTTLYA	26
GPTGSGKTTTLYAAL	26
LLISGPTGSGKTTTL	26
TGPTGSGKTTTLYAA	25
IFTGPTGSGKTTTLY	25
ILATGPTGSGKTTTL	25
ILVTGPTGSGKTTTL	25
ISGPTGSGKTTTLYA	25
IFVTGPTGSGKTTTL	24
GPTGSGKTTTLYSLL	24
LVTGPTGSGKTTTLY	24
ILVTGPTGSGKSTTL	23
LIFTGPTGSGKTTTL	23
FTGPTGSGKTTTLYS	23
LATGPTGSGKTTTLY	23
GPTGSGKTTTLYSVL	23
LISGPTGSGKTTTLY	23
GPTGSGKTTTLYATL	23
FVTGPTGSGKTTTLY	22
TGPTGSGKTTTLYSL	22
TGPTGSGKTTTLYSL	22
IILVTGPTGSGKTTT	22
TGPTGSGKTTTLYSV	22
GIIFVTGPTGSGKTT	21
IIFVTGPTGSGKTTT	21
ILVTGPTGSGKTVSL	21
GPTGSGKTTTLYSLV	21
GIILATGPTGSGKTT	21
IILATGPTGSGKTTT	21

B

VTGPTGSGKTTTLYA pattern pairs
VTGPTGSGKSTTLYG
ITGPTGAGKSTTDYE
FTGPVGSGKTQTMDD
VGGATGSGKSTTIYS
FVGGTASGKTTSLNA
ITGPMGSGKSSLVYA
VTGPTGSGKTTTLYA
VVGRSGSGKTTLINA
VTGPTGSGKTVSLYT
FTGPTGSGKTTTLYS
VTGGTGAGKTTLLKA
ATGPTGSGKTTTLYS
ITGPTGSGKSTTLYS
VTGPTGSGKTTTLYS
FTGPTGCGKSTTLYS
LSGPMGSGKTTTMYE
FAGGTASGKTTSLNA
VTGPTGSGKSISLYT
ISGPTGSGKTTTLYA
ITGPTGSGKSSTLYA
ICGETASGKTTTLNA
ISGGTGSGKTTLLNA
VNGETASGKTTTLMG
FSGPTGAGKTTLLNS
FAGPTNSGKTTSLTG
VTGPTGSGKSTTLYA

Table H.10 T2SS-4-15 patterns. A. Highly abundant pattern hits in T2SS-4-15. B. The most abundant pattern and its pairs in T2SS-4-15.

A

Patterns	Counts
QRGFTLLEIMVVIVI	16
RARGFTLLEVLVALA	16
QRGFTLLELLVVLVL	15
GFTLLEVLVALAIFA	15
GFTLLEVMVALAIFA	15
RRQRGFTLLEIMIAL	15
MLQRGFTLLELLVVL	14
GFTLLELLVVLVLVG	14
YRQRGFTLLEIMVVI	14
RQRGFTLLEIMVVIV	14
ARGFTLLEVLVALAI	14
RGFTLLEVLVALAIF	14
GFTLLELLVVLLIIA	14
ILVTGPTGSGKTTTL	13
ILVAGGTGSGKTTTL	13
LVAGGTGSGKTTTLN	13
VAGGTGSGKTTTLNS	13
RGFTLLEMLVVLVIA	13
VLISGGTGSGKTTTL	13
RGFTLLEIMVVIVIL	13
KRARGFTLLEVLVAL	13
ALRQRPDYIVMGEIR	13
RQRGFTLLEIMIALT	13
ILVTGPTGSGKTTTL	13
ILVTGPTGSGKTTTL	13

B

QRGFTLLEIMVVIVI pattern pair
QRGFTLLELLVVLVL
ERGFTLLEIMLVIFL
ARGFTLLEMLVVLVI
IRGFTLLEIMLVLLL
QRGFTLLEIMVVIVI
QRGFTLLEIMLVVLL
QQGFTLLEMMLVVLL
ARGFTLLEVLVALAI
QQGFTLLEMILAIAI
QAGFTLIEVMVAIML
PKGFTLLEVMVALAI
QRGFTLLEMMLVLLL
RRGFTLIELLVVLGI
NRGFTLIELLVVMAI
SCGFTLLELLVVLLI

Table H.11 T2SS-5-15 patterns. A. Highly abundant pattern hits in T2SS-5-15. B. The most abundant pattern and its pairs in T2SS-5-15.

A

Patterns	Counts
ALRQRPDYIVMGEIR	21
ALRQRPDYIVMGEIR	21
TGPTGSGKSTSLYAL	21
GPTGSGKSTSLYALL	21
ALRRDPDVLMVGEIR	20
IFVSGGTGSGKTTTL	20
ILVTGPTGSGKTSTL	20
LVTGPTGSGKTSTLY	20
VIVSGGTGSGKTTTL	20
IVSGGTGSGKTTTLN	20
VSGGTGSGKTTTLNA	20
LVTGPTGSGKSTTLY	19
VTGPTGSGKSTTLYA	19
GFFLVTGPTGSGKTT	19
FLVTGPTGSGKTTTL	19
LLRQDPDIIMVGEIR	19
FVSGGTGSGKTTTLN	19
TLIAGGTGSGKTTTL	19
LIAGGTGSGKTTTLN	19
VTGPTGSGKTSTLYA	19
NVIVSGGTGSGKTTT	19
SGGTGSGKTTTLNAL	19
GGTGSGKTTTLNALS	19
ILVTGPTGSGKSTTL	19
FLRQDPDIIMVGEIR	19
LRQRPDYIVMGEIRG	19
LRQRPDYIVMGEIRG	19
LLVSGGTGSGKTTLL	19
LVSGGTGSGKTTLLN	19

B

ALRQRPDYIVMGEIR pattern pairs
ILRQDPDIIMIGEMR
ALRMRPDRLIVGEVR
ALRRDPDVLMVGEIR
FLRQDPDVIMVGEIR
LLRQDPDIIMVGEIR
SLRMRPDRVVIGEVR
ILRQDPDVILIGEIR
ALRMRPDRILVGEVR
LLRQDPDIVMVGEIR
MLRQDPDIIMIGEIR
LLRQDPDIILVGETR
ALRMRPDRIIIGEVR
ALRMRPDRIIVGETR
FLRQDPDIIMVGEIR
ALRQRPDYIVMGEIR
ALRQRPDYIVMGEIR
ALRHRPEYLLVGEVR
ALRMRPDRIVVGECR
AMRLDPDAILNGEIR
ILRQDPDIIMIGEIR
ILRQDPDVIMVGEMR

Table H.12 T2SS-total-15 patterns. A. Highly abundant pattern hits in whole T2SS. B. The most abundant pattern and its pairs in whole T2SS data

A		B
Patterns	Hits	Few ILVAGGTGSGKTTTL pattern pairs
ILVAGGTGSGKTTTL	105	VIISGGTGSGKTTLL
LLVTGPTGSGKTTTL	104	VLVTGPTGSGKSTTL
ILVTGPTGSGKTTTL	104	ILVTGPTGSGKSTTL
ILVTGPTGSGKTTTL	104	LLVTGPTGSGKTTTL
LLISGPTGSGKTTTL	104	ILVGGATGSGKSTTI
ILVTGPTGSGKTTTL	104	FIISGGTGSGKTTTL
ILVTGPTGSGKTTTL	104	MLFVGGTASGKTTSL
LLVTGPTGSGKTTTL	104	ILISGETGTGKTEVL
ILVTGPTGSGKTTTL	104	ILVTGPTGSGKSTTL
ILVSGPTGSGKTTLL	103	ILVTGPTGSGKTTTL
VLVTGPTGSGKTTTL	103	ILVAGGTGSGKTTTL
VLVTGPTGSGKTTTL	103	IFVTGPTGSGKTTTL
VLVSGPTGSGKTTTL	102	LLVTGATGSGKSTTL
ILATGPTGSGKTTTL	102	YVVAGTTGSGKSTTL
LVAGGTGSGKTTTLN	101	IIVTGPTGSGKSTTL
VLLTGPTGSGKTTTL	101	ILVTGPTGSGKTVSL
VLISGGTGSGKTTTL	101	ILVVGRSGSGKTTLI
LVSGPTGSGKTTLLN	101	YLIVGSTGSGKTTFL
FLVTGPTGSGKTTTL	101	VLVSGPTGSGKTTTL
VTGPTGSGKTTTLYA	100	VLVTGPTGSGKTLSL
GPTGSGKTTTLYAML	100	AIIAGGTGSGKTTTL
VTGPTGSGKTTTLYA	100	VLLTGPTGSGKTTTL
VTGPTGSGKTTTLYA	100	ILVTGPTGSGKTVSL
LISGGTGSGKTTTLN	99	FIMAGQTGSGKTTTI
LVTGPTGSGKTTTLY	98	LIFTGPTGSGKTTTL
LVTGPTGSGKTTTLY	98	VLVTGGTGAGKTTLL
NILVAGGTGSGKTTT	98	TLIVGGTGSGKTTTL
GPTGSGKTTTLYAAL	98	ILATGPTGSGKTTTL
LVSGPTGSGKTTTLY	98	ILITGPTGSGKSTTL

H.3 Type 3 Secretion Related Patterns

Table H.13T3SS-1-15 patterns. A. Highly abundant pattern hits in T3SS-1-15. B. The most abundant pattern and its pairs in T3SS-1-15.

A		B
Patterns	Counts	LAKIPYLGDIPILGA pattern pairs
VPLLGDIPVIGALFR	5	LAKIPYLGDIPILGA
SKIPLLGDIPFIGSL	5	ADKVPLLGDIPVIGA
RKIPLLGDIPILGRL	5	ERKIPLLGDIPILGR
PYLGDIPILGALFSK	5	NDKIPLLGDIPLAGR
PLLGDIPVIGALFRS	5	NSKIPLLGDIPFIGS
PLLGDIPLAGRLFQF	5	AKIPYLGDIPILGAL pattern pairs
PLLGDIPILGRLFKT	5	AKIPYLGDIPILGAL
PLLGDIPFIGSLFRS	5	DKVPLLGDIPVIGAL
NSKIPLLGDIPFIGS	5	RKIPLLGDIPILGRL
NNSKIPLLGDIPFIG	5	DKIPLLGDIPLAGRL
NDKIPLLGDIPLAGR	5	SKIPLLGDIPFIGSL
LLGDIPVIGALFRSD	5	KIPYLGDIPILGALF pattern pairs
LLGDIPILGRLFKTT	5	KIPYLGDIPILGALF
LGDIPVIGALFRSDS	5	KVPLLGDIPVIGALF
LAKIPYLGDIPILGA	5	KIPLLGDIPILGRLF
KVPLLGDIPVIGALF	5	KIPLLGDIPLAGRLF
KIPYLGDIPILGALF	5	KIPLLGDIPFIGSLF
KIPLLGDIPLAGRLF	5	IPYLGDIPILGALFS pattern pairs
KIPLLGDIPILGRLF	5	IPYLGDIPILGALFS
KIPLLGDIPFIGSLF	5	VPLLGDIPVIGALFR
IPYLGDIPILGALFS	5	IPLLGDIPILGRLFK
IPLLGDIPLAGRLFQ	5	IPLLGDIPLAGRLFQ
IPLLGDIPILGRLFK	5	IPLLGDIPFIGSLFR
IPLLGDIPFIGSLFR	5	PYLGDIPILGALFSK pattern pairs
HNDKIPLLGDIPLAG	5	PYLGDIPILGALFSK
GERKIPLLGDIPILG	5	PLLGDIPVIGALFRS
ERKIPLLGDIPILGR	5	PLLGDIPILGRLFKT
DKVPLLGDIPVIGAL	5	PLLGDIPLAGRLFQF
DKIPLLGDIPLAGRL	5	PLLGDIPFIGSLFRS
AKIPYLGDIPILGAL	5	

Table H.14 T3SS-2-15 patterns. A. Highly abundant pattern hits in T3SS-2-15. B. The most abundant pattern and its pairs in T3SS-2-15.

A

Patterns	Counts
VPAAVPAEVPAEVPA	7
PAAVPAEVPAEVPAE	6
PLLGDIPILGELFKS	5
YGIALAATLFVMAPV	4
VVLGLLRSALGIQQV	4
VTTILMAMGMSMVSP	4
VSVVLFLVRNALGTQ	4
VSNVLLALGMQMVSP	4
VQQVPPNMALYGIAL	4
VPPNMALYGIALAAT	4
VLTLLRSALGVQQAP	4
VLMLTRNAMGVQQVP	4
VLGLLRSALGIQQVP	4
VLFLVRNALGTQSIP	4
VIDLIVANVLTAMGM	4
VANVLTAMGMMMLSP	4
VANILLALGMQMISP	4
SVVLFLVRNALGTQS	4
SIVLTLLRSALGVQQ	4
RSALGIQQVPPNLVL	4
RNAMGVQQVPPNMAL	4
QQVPPNMALYGIALA	4
PPNMALYGIALAATL	4
PNMALYGIALAATLF	4
NILLALGMQMISPTT	4
NAMGVQQVPPNMALY	4
MQMISPTTISVPFKL	4
MLTRNAMGVQQVPPN	4
LYGIALAATLFVMAP	4

B

Highest Pairs
VPAAVPAEVPAEVPA pattern pairs
TPTAAPAEVPVESSA
FPADSPAAVPAAVPA
SPAAVPAAVPAEVPA
VPAAVPAEVPAEVPA
VPAEVPAEVPAEVPA
VPAEVPAEVPAEFPA
VPAEVPAEFPADRAG
PAAVPAEVPAEVPAE pattern pairs
PTAAPAEVPVESSAP
PADSPAAVPAAVPAE
PAAVPAAVPAEVPAE
PAAVPAEVPAEVPAE
PAEVPAEVPAEVPAE
PAEVPAEVPAEFPAD
PLLGDIPILGELFKS pattern pairs
PFIGDVPILGPLFRV
PLLGDIPGLGFLFSS
PILGDIPIVQYFFGS
PLLGDLPIIGAFFRN
PLLGDIPILGELFKS

Table H.15 T3SS-3-15 patterns. A. Highly abundant pattern hits in T3SS-3-15. B. The most abundant pattern and its pairs in T3SS-3-15.

A

Patterns	Counts
KVPLLGDLPLLGALF	9
KVPFLGDIPYLGRLF	9
VPFLGDIPYLGRLFR	8
PLLGDLPLLGALFRR	7
KVPGLGQLPLLGRLF	7
IPLLGDIPVVGHLFR	7
VPLLADIPLVGALFK	6
VPGLGQLPLLGRLFS	6
TKVPFLGDIPYLGRL	6
SVDKVPLLGDLPLLG	6
SIPFLGDIPGLGRLF	6
RVPLLADIPLVGALF	6
QTKVPFLGDIPYLGR	6
NKIPLLGDIPVVGHL	6
LLGDLPLLGALFRRS	6
KIPLLGDIPVVGHLF	6
KIPILGSIPFIGKLF	6
IPFLGDIPGLGRLFR	6
DKVPLLGDLPLLGAL	6
VDKVPLLGDLPLLGA	5
TNSVDKVPLLGDLPL	5
SQTKVPFLGDIPYLG	5
PLLGDIPVVGHLFRN	5
PFLGDIPYLGRLFRK	5
NSVDKVPLLGDLPLL	5
LPGIGELPVLGALFR	5
LLGDIPVVGHLFRND	5
IPILGSIPFIGKLFG	5
DRVPLLADIPLVGAL	5
ALPGIGELPVLGALF	5

B

Highest Pairs
KVPLLGDLPLLGALF pattern pairs
KIPLLGDIPVVGHLF
KVPLLGDLPLLGALF
RVPLLADIPLVGALF
SIPFLGDIPGLGRLF
KVPFLGDIPYLGRLF
KVPFLGDLPIIGTFF
ALPGIGELPVLGALF
KVPGLGQLPLLGRLF
GLPWLSELPLIGALF
VPLLGDLPLLGALFR pattern pairs
IPLLGDIPVVGHLFR
VPLLGDLPLLGALFR
VPLLADIPLVGALFK
VPLVQDVPLARALYR
IPFLGDIPGLGRLFR
VPFLGDIPYLGRLFR
VPFLGDLPIIGTFFK
LPGIGELPVLGALFR
VPGLGQLPLLGRLFS
LPWLSELPLIGALFG

Table H.16 T3SS-4-15 patterns. A. Highly abundant pattern hits in T3SS-4-15. B. The most abundant pattern and its pairs in T3SS-4-15.

A

Patterns	Counts
EKTERPTPKRLRDSR	6
DIVVVNPTHIAVAIY	6
VVVNPTHIAVAIYLD	5
VVNPTHYAVALAYEP	5
VVNPTHIAVAIYLDP	5
VNPTHFAVGLYYRPG	5
VMVVNPTHYAVALAY	5
VLVTNPTHLAVALYY	5
VKQEYKEMEGDPHIK	5
TEKPTEKKLRDARKD	5
SEKTEQPTEKKLRDG	5
SEEKTEKPTEKKLRD	5
RMSKDEVKQEYKEME	5
MVVNPTHYAVALAYE	5
MSKDEVKQEYKEMEG	5
MSEKTEQPTEKKLRD	5
MAEKTEKPTSKKLKD	5
LVTNPTHLAVALYYA	5
LVNPTHFAVGLYYRP	5
LRMSKDEVKQEYKEM	5
KTEQPTEKKLRDGRK	5
KTEKPTEKKLRDARK	5
KQEYKEMEGDPHIKQ	5
KGLRMSKDEVKQEYK	5
IVVVNPTHIAVAIYL	5
GLRMSKDEVKQEYKE	5
GEKTERPTPKRLRDS	5
EYKDNEGDPHLKSAR	5
EKTEQPTEKKLRDGR	5
VPLLGDIPVLGRLFR	4

B

Highest Pairs
EKTERPTPKRLRDSR pattern pairs
EKTEQPTEKKLRDGR
EKTERPTPKRLRDSR
EKTEDATPQKLQESR
EKTEKPTEKKLRDAR
EKTEKPTEKKIKDSA
EKTEKPTSKKLKDES
DIVVVNPTHIAVAIY pattern pairs
VAVVRNPTHIAVCLG
TVLVTNPTHLAVALY
DVLLVNPTHFAVGLY
KVMVVNPTHYAVALA
TFVMANPTHIAMLIY
DIVVVNPTHIAVAIY
SEKTEQPTEKKLRDG pattern pairs
SEKTEQPTEKKLRDG
GEKTERPTPKRLRDS
EEKTEKPTEKKLRDA
AEKTEKPTEKKIKDS
AEKTEKPTSKKLKDE
EKTEQPTEKKLRDGR pattern pairs
EKTEQPTEKKLRDGR
EKTERPTPKRLRDS
EKTEKPTEKKLRDAR
EKTEKPTEKKIKDSA
EKTEKPTSKKLKDES

Table H.17 T3SS-5-15 patterns. A. Highly abundant pattern hits in T3SS-5-15. B. The most abundant pattern and its pairs in T3SS-5-15.

A

Patterns	Counts
SKVPLLGDIPVLGHL	6
KVPLLGDIPVLGHLF	6
VSKVPFLGDVPALGH	5
VPLLGDIPYLGALFR	5
VPLLGDIPVLGHLFK	5
VPFLGDVPALGHLFR	5
SKVPFLGDVPALGHL	5
SESKVPLLGDIPVLG	5
RKVPLLGDIPYLGAL	5
QVSKVPFLGDVPALG	5
PLLGDIPYLGALFRS	5
PLLGDIPVLGHLFKS	5
PFLGDVPALGHLFRN	5
LLGDIPVLGHLFKST	5
KVPLLGDIPYLGALF	5
KVPFLGDVPALGHLF	5
ESKVPLLGDIPVLGH	5
YGELVEVDDKLGVEI	4
WVATAVGELIDNQRG	4
VVSRPVILTQENIPA	4
VSRPVILTQENIPAI	4
VSRPILLTQENTPAI	4
VSFSSFPALLLFTTL	4
VSEGGSLLIGGYTRE	4
VPQGKSLLIGGYTHE	4
VPMLGDMPILGNLFR	4
VPKGSSLLVGGYSRD	4
VKQGQSLLIGGVYRD	4
VATAVGELIDNQRGA	4

B

SKVPLLGDIPVLGHL pattern pairs

RKVPLLGDIPYLGAL
SKVPFLGDVPALGHL
SKVPLLGDIPVLGHL
WGIPLLRDIPFLGRL
DKVPMLGDMPILGNL
SKLPFLGDLPVIGQF

VSKVPFLGDVPALGH pattern pairs

LRKVPLLGDIPYLGA
VSKVPFLGDVPALGH
ESKVPLLGDIPVLGH
MDKVPMLGDMPILGN
ESKLPFLGDLPVIGQ

QVSKVPFLGDVPALG pattern pairs

QLRKVPLLGDIPYLG
QVSKVPFLGDVPALG
SESKVPLLGDIPVLG
DMDKVPMLGDMPILG
TESKLPFLGDLPVIG

Table H.18 T3SS-total-15 patterns. A. Highly abundant pattern hits in whole T3SS. B. The most abundant pattern and its pairs in whole T3SS data

A

Patterns	Counts
VPLLGDLPLLGALFR	32
VPLLGDIPVIGALFR	30
VPLLGDIPVLGRLFR	29
KVPLLGDLPLLGALF	29
KVPFLGDIPYLGRLF	29
VPFLGDIPYLGRLFR	28
PLLGDLPLLGALFRR	28
PLLGDIPVLGRLFRR	28
VPLLGDIPYLGALFR	27
PLLGDIPILGELFKS	27
KVPLLGDIPVIGALF	27
KVPFLGDIPFLGNLF	27
KIPLLGDIPILGRLF	27
IPLLGDIPILGRLFK	27
VPFLGDIPFLGNLFK	26
PLLGDIPYLGALFRS	26
PLLGDIPVIGALFRS	26
PLLGDIPILGRLFKT	26
KVPLLGDIPVLGHLF	26
KIPLLGDIPLAGRLF	26
KIPLLGDIPFIGSLF	26
KFPLLGDIPILGELF	26
IPFLGDIPGLGRLFR	26
VPLLGDIPVLGHLFK	25
SVPLLGDIPVLGRLF	25
SIPFLGDIPGLGRLF	25
PLLGDIPVLGHLFKS	25
PFLGDIPFLGNLFKK	25
KVPLLGDIPYLGALF	25
KVPGLGQLPLLGRLF	25
KIPYLGDIPILGALF	25
KIPLLGSIPYLGRLF	25

B

VPLLGDLPLLGALFR pattern match:
IPLLGDIPVVGHLFR
VPLLGDIPYLGALFR
VPFLGDVPALGHLFR
VPLLGDLPLLGALFR
IPYLGDIPILGALFS
VPLLGDIPVLGHLFK
VPLLADIPLVGALFK
LPGLGSIPLIGGLFR
IPLLRDIPFLGRLFD
LPFIGDVPILGPLFR
VPLLGDIPGLGFLFS
VPLVQDVPLARALYR
IPFLGDIPGLGRLFR
VPFLGDIPYLGRLFR
VPLLGDIPVLGRLFR
VPLLGDIPVIGALFR
VPFLGDIPFLGNLFK
IPLLGDIPILGRLFK
VPMLGDMPILGNLFR
VPFLGDLPIIGTFFK
VPMLSKLPLVGALFR
LPGIGELPVLGALFR
VPGLGQLPLLGRLFS
IPLLGSIPYLGRLFS
IPLLGDIPLAGRLFQ
LPFLGDLPVIGQFFR
VPGLSKLPLIGWLFR
IPLLGDIPFIGSLFR
LPWLSELPLIGALFG
VPLLGDLPIIGAFFR
VPFLSKLPVVGALFG
FPLLGDIPILGELFK

H.4 Type 4 Secretion Related Patterns

Table H.19 T4SS-1-15 patterns. A: Highly abundant pattern hits in T4SS-1-15. B. The most abundant pattern and its pairs in T4SS-1-15.

A

Patterns	Counts
AALGAALGAA	26
FRYDAIGRLV	21
YRYDGLGRLV	19
TRYRYDAAGR	18
YRYDAAGRPV	17
RYDAIGRLVE	17
GDPDRPVIVG	16
VPRAGEEVVV	15
GDPDRPLIVG	15

B

AALGAALGAA Pattern Pairs
AVIGAAVGAA
AAVGAAATVA
AALYAALSAY
FKLMAALGAW
AALGAWFGVI
PAKPALLGAA
ALLGAAVGYV
FKLLAALGAW
AALGAWCGLK
IKLFAALGAW
AALGAWFGLN
LVGGAILGAA
AAIGAAVQAA
KIGGAALGAA
AALGAALGAA
AALGAAPSLL
AVLGAGLAAI
LKLLAALGAW
AALGAWLGWE
FKLIAALGAW
AALGAWLGIS
AALGGWVGAL
GALGVAINAA
AELGAVLHAF
AAVQAQLGAA
AMLLAGFGAA

Table H.20 T4SS-2-15 patterns. .A. Highly abundant pattern hits in T4SS-2-15. B. The most abundant pattern and its pairs in T4SS-2-15.

A		B
Patterns	Counts	RYAYDAAGRL pattern pairs
RYAYDAAGRL	34	HYEYDALGRR
YEYDAYGRLT	32	RYRYDEAGRL
YAYDAAGRLV	32	RYVYDRLGHL
YDYDAAGRMV	26	RYQPDAAGRL
YRYDVLGQLT	23	AYEYDAYGRL
LGYDALGRLT	22	KYRYDAADRL
RYRYDEAGRL	21	RYAYDLGRRM
AYEYDAYGRL	20	RFAWDAGDRL
TRYTYDLAGR	19	RYAYDALDNV
RYGYDALGNL	18	RYRYDGFGRM
TRYAYDAAGR	18	RYTYDLAGRV
RFSYDKAGRL	18	TFAYDVRGRL
YDALGRLTGL	18	RYGYDALGNL
YRYDEAGRLN	17	LYVYDDAGRV
FSYDALGQLI	16	RYTRDAFGRA
RYTYDLAGRV	16	RFTYDIHGRL
WSYLYDALGR	16	SYLYDALGRR
FSYDKAGRLV	16	RYDYDARGRL
YRYDLLGRRT	16	RFEYDDAGRV
YEYDALGRRT	15	RFAYDPAQRV
LERDAAGRLT	15	HYTYDPAGRL
YLYDALGRRT	15	RYAYDAAGRL
RYDYDARGRL	15	RYQLDAAGRV
YTYDPAGRLA	15	RFAYDSADRL
YQLDAAGRVT	15	RFSYDKAGRL
YEYDAHGRLV	15	RYRYDRFGRM
RYRYDVLGQL	15	RYRYDVLGQL

Table H.21 T4SS-3-15 patterns. A. Highly abundant pattern hits in T4SS-3-15. B. The most abundant pattern and its pairs in T4SS-3-15.

A

Patterns	Counts
YEYDAAGNLV	22
TDPLGRTTRY	20
TDPLGATTRY	18
YEYDPAGRLI	17
YTYDAFGRRL	16
YHYDALGRRT	16
TDPLGRTTTV	16
VTDPLGATTR	16
WRYGYDALGR	16
GFTLVELLVV	16
RGFTLIELVV	15
GFTLIELVVV	15
KGFTLIELMI	15
GFTLIELMIV	15
YRFTYDAVGR	15
TDPAGAVTRY	15
SVTDPLGATT	15
VTDPLGATTH	15
YAYDPAGRLV	15
YEYDAADRLI	15
VTDPMGATTR	15
TDPMGATTRL	15

B

YEYDAAGNLV pattern pairs
ARYDAKGNLV
YEYEANGQLL
YKYDAWGNLI
YEYNAAHQLI
YERDAAGQII
YEWDALSNLT
FSYDAADNLA
MKYDSWGNLV
MEYDAAGRVI
YAYDAAGDLA
WEWDGEGNLV
YEYDPAGRLI
YELNAAGDLV
YTYDAAGDLV
YEHDAAGRLI
YNYDQAGNVL
YEWDDRGNLL
YEYDLHGQLL
YEYDAADRLI
YAYDLSGNVV
YAFDRVGNLV
FEYDTAGRVV
YEYDAAGNLV
YYYDKAGNLL

Table H.22 T4SS-4-15 patterns. A. Highly abundant pattern hits in T4SS-4-15. B. The most abundant pattern and its pairs in T4SS-4-15.

A

Patterns	Counts
AAAAAAAVAV	20
AAAAAAVAVA	20
VTYDANGNVT	20
AVAAAAAAAV	18
VAAAAAAAVA	18
GNQTTYTYDA	18
SYDANGNLTK	17
YRYDAVGRLT	16
FAVAAAAAAA	15

B

AAAAAAAVAV pattern pairs
AIAIAAANKV
LAALAAAGAI
AAAALVADIV
LAAASGAVIV
ALVAAAVFAV
VAAAVFAVAA
FAVAAAAAAA
AVAAAAAAAV
VAAAAAAAVA
AAAAAAAVAV
AAAAAAVAVA
AAAAAVAVAV
AAAVAVAVIG
ARSAAAALAV
SAAAALAVAS
LLAAAAAGAT
AAAAAGATAV
DAAAKAAAAL
AAAKAAAALT
AALVGAQVAV

Table H.23 T4SS-5-15 patterns. A. Highly abundant pattern hits in T4SS-5-15. B. The most abundant pattern and its pairs in the T4SS-5-15.

A

Pattern	Counts
YDALGRVTEV	21
LAYDAMGRLT	18
LIELMIVIAI	17
YQYDAAGNLT	17
YDAMGRLTDV	17
QGFTLIELMI	16
TLIELMIVVA	16
LIELMIVVAI	16
QGFTLIELVI	16
GFTLIELVIV	16
KGFTLIELMI	16
TLIELMIVVA	16
LIELMIVVAI	16
KGFTLIELMI	16
TLIELMIVVA	16
LIELMIVVAI	16
KGFTLIELMI	16
TLIELMIVIA	16
KGFTLIELMI	16
TLIELMIVVA	16
LIELMIVVAI	16
TLIELMIVVA	16
LIELMIVVAI	16
GFTLIELMIV	15
GFTLIELMIV	15
GFTLIELMIV	15
GFTLIELMIV	15
GFTLIELMIV	15
TYGYDAMGRL	15
YGYDAMGRLN	15
GFTLLELLIV	15
GGFTLIELMI	15
GFTLIELMIV	15
KGFTLIELIV	15

B

YDALGRVTEV pattern pairs
YDAADRVTAV
HDALGRVTSM
YDAFDRLREV
YDAQGRLTLA
SDALGTATEL
TDAAGRTTEY
YDILGRLLEV
YDAFGRRTEK
VEALGYETEV
LDAAGRVKSV
YDNGGRLTEA
YDALDRLTQV
YDILGRPTTV
YDAMGRLTDV
YDANGNLTDV
LDAIGQVTAV
YDASGNLTEA
YDALGRKAEV
RDARGRITEL
HDYLGRLTEA
VDAAGNVTET
YDALGRVTEV
YDALGRKTAL

Table H.24 T4SS-total-15 patterns. A. Highly abundant pattern hits in whole T4SS. B. The most abundant pattern and its pairs in whole T4SS data

Pairs	Counts	Some GFTLVELLVVIAIIG pattern pairs
GFTLVELLVVIAIIG	46	GFTLIEVLVSIAIFA
GFTLVELLVVIAIIG	46	GFTLIELIMCIIILG
GFTLIELMIVVAIIG	44	GFTLIELMIVVAIIG
GFTLIELMIVVAIIG	44	GFTLIELVVVIVILG
GFTLVELMVVIVIIG	44	GVTLIELLIVIVVLG
GFTLIELMIVVAIIG	44	GFTLIELMIVVAVIG
GFTLIELMIVVAIIG	44	GFTLIELVVVVAVLG
GFTLIELMIVVAIIG	44	GMTLLELLIVVTLIG
GFTLIELMIVVAIIG	44	GFTLIELMIVVAIIG
QKGFTLIELMIVIAI	44	GFTLIELVIVIVVLG
KGFTLIELMIVIAIV	44	GFTLIELVTVVVILG
GFTLIELMIVVAIIG	44	GFTLIELMVTIAVLA
GFTLIELMIVVAIIG	44	GFTLVELMVVIVIIG
GFTLIELMIVVAIIG	44	GFTLVEMMIVLMIIS
GFTLIELLVVISIIG	44	GFTLIELMIVVAIIG
GFTLIELVVVIVILG	43	GFTLIELMIVVAIIG
QKGFTLIELMIVVAI	43	GFTLIELMIVVAIIG
KGFTLIELMIVVAII	43	GFTLIELMIVVAIIG
QKGFTLIELMIVVAI	43	GFTLIELMIVIAIVG
KGFTLIELMIVVAII	43	GFTLIELMIVVAIIG
QKGFTLIELMIVVAI	43	GITLIELMIVVAIIG
KGFTLIELMIVVAII	43	GFTLIELMIVVAIIG
QKGFTLIELMIVVAI	43	GFTLIEVMVVIVILG
KGFTLIELMIVVAII	43	GFTLVELLVAITIFV
QKGFTLIELMIVVAI	43	SITLVEMMVVITLIG
KGFTLIELMIVVAII	43	GFTLIELMVVIGIIA
GFTLIEVMVVIVILG	43	GFTLIELMVTIAVLA
GFTLIELLVVIIIIG	43	GFSLLELMVALAIFG
RGFTLIELMIVVAII	42	GFTIIELLVTLAILA
GFTLIELVIVIVVLG	42	GFTLLEMLVVLVIAG

H.5 Type 5 Secretion Related Patterns

Table H.25 T5SS-1-10 patterns. A. Highly abundant pattern hits in T5SS-1-10. B. Sample repeat pairs and its pairs in the T5SS-1-10.

A

Pairs	Counts
IVAGAYDYTL	16
IVAGAYDYTL	16
VDAGAYEYRL	15
VNAGGAGALT	14
VAGAYDYTLA	14
LNVTGNATGA	14
AGAYDYYLYK	14
VAGAYDYTLA	14
SLEAGYGFAL	14
TDRLVITGDT	13
GYSAGLYATW	13
IFAGAYEYSL	13
GYSAGLYATW	13
TVSLEAGYGF	13
TASVEAGYSF	13
FEPQVQFIYQ	12
SLEAGLGFDA	12
LNVTGNANGN	12
GYSVGLYGTW	12
LTVSTLSGTG	12
FNTVLGDDSS	12
GYSVGLYGTW	12
GYSAGLYGTW	12
LTASLEAGYT	12
LTVQGNYVGN	12
TLSAEAGYPF	12
GYSVGLYGTW	12
VTVNNAGGTG	12
TVNNAGGTGA	12

B

IVAGAYDYTL pattern pairs
TIAGVYDTTL
AVGGAYEYFL
IVAGAYDYTL
IVGGAYDYYL
VEAGAWVYTL
IVAGGYDYDV
VSAGAFDYRL
VSAGAYDYYL
IVAGAYDYDV
IVAGAYDYTL
IFAGAYEYSL
VQAGAYEYIL
VDAGAYEYRL
VVAGAYGYRL
IVAGAYDYNV
AVAGPYEYRL
LTVSTLSGTG pattern pairs
LTMASLNGTG
VTVSQLRATG
LTVSTLSGTG
LALSSLSGNG
ITVTSLGGTG
LTASTLSGNG
QTVTTVSGTL
LTTNELSGSG
LNISNLSGNG
TTLSSLSGTG
VTVIQNSGTG
LTRSQLSGQN

Table H.26 T5SS-2-10 patterns. A. Highly abundant pattern hits in T5SS-2-10. B. Sample repeat pairs and its pairs in the T5SS-2-10.

A

Patterns	Counts
IEPQAQLVYQ	14
EPQAQLVYQQ	13
VFEPQAQLVY	13
FEPQAQLVYQ	13
AGGSGAVASG	12
IEPQAQVIYQ	12
EPQLQVVYQS	11
PQAQVIYQYL	11
PQAQLVYQQL	11
EPQAQLVYQR	11
VDAGAYEYRL	11
PQAQVIWQNL	10
KIEPQAQVIY	10
QIEPQAQLVY	10
PQAQLVYQRL	10
GATNVAAGTL	10
LLIHGNVSGK	9
PQIQLIYQHL	9
LEPQLQVVYQ	9
PQLQVVYQSL	9
GAYLDGWLQY	9
PQVQVVYQHL	9
NAQGTGSLAL	9
ATATGAQSVA	9
TLTVTGDYAG	9
AGASGAIAIG	9
LTGNGSVAIG	9
LKIEPQAQVI	9
EPQAQVIYQY	9

B

IEPQAQLVYQ pattern pairs
FDPQIQLIYQ
LEPQLQVVYQ
VEPQMELVYG
FDPQVQVVYQ
IEPQAQVIWQ
IEPQAQVIYQ
IEPQTQVIYQ
MEPQAQVIGQ
IEPQAQLVYQ
FEPQAQLVYQ
IIPQAQLVYN
ITPQVQLQYS
VEPQLQLVHQ
IEPQAELVYG

Table H.27 T5SS-3-10 patterns. A. Highly abundant pattern hits in T5SS-3-10. B. Sample repeat pairs and its pairs in the T5SS-3-10.

A

Pairs	Counts
LEPQAQLVYQ	20
EPQAQLVYQR	19
PQAQLVYQRL	18
PQAQLIYQRL	18
LEPQAQLIYQ	18
IEPQAQLVYQ	18
EPQAQLIYQH	18
EPQAQLAYQR	18
SGTGTLTLTN	17
PQAQLAYQRL	17
FVPQAQLIYQ	17
FEPQAQLAYQ	17
VFEPQAQLAY	16
TLEPQAQLIY	16
PQVQVVYQHL	16
PQAQLIYQHT	16
PQAQIAYQRL	16
LGSGTLTLNG	16
ILEPQAQLVY	16
FIEPQAQLVY	16
FEPQAQVAYQ	16
FEPQAQIAYQ	16
EPQAQLVYQN	16
EPQAQIAYQR	16
VFVPQAQLIY	15
TLTINGDYSG	15
SGSGTLTLSG	15
GYFIEPQAQL	15
WTLEPQAQLI	14

B

LEPQAQLVYQ pattern pairs:
LQPELQLVYT
LEPQAQVIVS
MEPQAQVIGQ
VEPQAQLIAQ
LEPQVQIIHQ
VEPQGQLKYQ
LTPQAQVVWQ
IEPQAQVIYQ
IEPQGQIIYQ
LEPQAQVIYQ
VEPQVQAVYN
LEPQLQYQVQ
LTPQAQIIWQ
IEPQAELVYG
VEPQAELVWG
IEPQTQLTYS
LEPQLQYTWQ
IEPQTQVIYQ
IEPQAQVIYS
VEPQVQLSYM
IEPQAQLSYL
IEPQAQVIWQ
IEPQAQVIWQ
VEPQLQLVHQ
VEPQLQLIHQ
LEPQLQLTHQ
LEPQAQLVWQ
VEPQAQLMYQ
IEPQAQLAYQ

Table H.28 T5SS-4-10 patterns. A. Highly abundant pattern hits in T5SS-4-10. B. Sample repeat pairs and its pairs in the T5SS-4-10.

A

Pairs	Counts
EPQAQLMYQY	14
PQAQVVYQYL	13
FEPQMQLIYQ	13
GAGTLTLSGN	13
GTLTLSGNSA	13
IEPQAQLAYQ	13
VEPQAQLMYQ	13
PQAQLMYQYL	13
AGASTLTLSG	12
GASTLTLSGT	12
QAQVVYQYLQ	12
EPQMQLIYQY	12
PQMQLIYQYL	12
IEPQAQVIWQ	12
PQAQVIWQNL	12
GDVTLGSGTL	12
AGTLTLSGNS	12
SGAGTLVLTA	12
LTLTGNSTYT	12
IEPQAQLIYQ	12
PQAQLIYQHL	12
AFGANAAAGG	12
ATAAGSNATA	12
FIEPQAQLAY	12
EPQAQLAYQH	12
PQAQLAYQHL	12
GTVTATGGTL	12
GFYMDGVLSY	11
NPQAQVVYQY	11

B

EPQAQLMYQY pattern pairs
DPQVQIVYQY
NPQAQVVYQY
EPQMQLIYQY
EPQMQLVYQY
EPQMQLVYQY
EPQVQVIYSY
EPQAQVIWQN
TPQAQLMWSK
EPQAQLIAQK
EPQGQLKYQY
EPQAQLIYQH
EPQAQLAYQH
EPQAQVSMLY
EPQAQLMYQY

Table H.29 T5SS-5-10 patterns. A. Highly abundant pattern hits in T5SS-5-10. B. Sample repeat pairs and its pairs in the T5SS-5-10.

A

Pairs	Count
AIGTAANASG	13
GYSTGLYGTW	12
LGDDQSATDK	12
AGTGNLTVTN	11
SGTGNVAVSG	11
SGTTTVAGGT	11
AATASGVDSV	11
SGYSVGLYGT	11
GYSVGLYGTW	11
GVYLDSWAQY	11
INTVLGDDTS	11
TTGGSITTTG	11
TASLEAGYRY	11
SVALGASATA	11
MNKIYRVVWN	11
GDDQSATDKL	11
SGYSAGVYGT	11
GTLTLANSGA	11
GNTTVSGGSL	11
ATLSNAGTVA	11
AATGAGATAT	11
GTGNLTVTNN	10
TLTVTGTSSI	10
ASLDNAGTVS	10
GTTTLTGAST	10
TLNNAGTVAL	10
VGTGTLTFSG	10
GTGTLTFSGN	10
GTLALTGAGS	10

B

AIGTAANASG pattern pairs
ALGTASNATG
LDGAAASASG
AVGQAAVAPG
FSGTMANASG
AVGYAATASA
NFVTIANASG
APGTLLDASG
AIGTAANASG
AIGSYANAVL
AAGTVRIASG
AVGVLARASG
ASGTNASATG
AIGNVAAATG

Table H.30 T5SS-total-10 patterns. A. Highly abundant pattern hits in whole T5SS. B. The most abundant pattern and its pairs in whole T5SS data

A		B
Pairs	Counts	LEPQAQLVYQ pattern pairs
LEPQAQLVYQ	72	LQPELQLVYT
IEPQAQLVYQ	67	LEPQAQVIVS
IEPQAQLVYQ	67	MEPQAQVIGQ
FEPQAQLVYQ	65	VEPQAQLIAQ
LEPQAQLIYQ	64	LEPQVQIIHQ
IEPQAQLIYQ	63	VEPQGQLKYQ
FEPQAQLAYQ	62	LTPQAQVVWQ
IEPQAQVIYQ	61	IEPQAQVIYQ
LEPQAQVIYQ	61	IEPQGQIIYQ
PQAQLIYQRL	61	LEPQAQVIYQ
EPQAQLVYQR	60	VEPQVQAVYN
EPQAQLVYQR	60	LEPQLQYQVQ
EPQAQLVYQN	60	LTPQAQIIWQ
EPQAQLVYQQ	59	IEPQAELVYG
EPQAQLIYQH	59	VEPQAELVWG
EPQAQLIYQH	59	IEPQTQLTYS
VFEPQAQLVY	59	LEPQLQYTWQ
VEPQAQLMYQ	58	IEPQTQVIYQ
IEPQAQLAYQ	58	IEPQAQVIYS
VIEPQAQLIY	58	VEPQVQLSYM
PQAQLIYQHL	58	IEPQAQLSYL
PQAQLVYQRL	58	IEPQAQVIWQ
PQAQLVYQRL	58	IEPQAQVIWQ
FEPQAQVIYQ	58	VEPQLQLVHQ
VDAGAYEYRL	57	VEPQLQLIHQ
VDAGAYEYRL	57	LEPQLQLTHQ
VDAGAYEYRL	57	LEPQAQLVWQ
FEPQMQLIYQ	57	VEPQAQLMYQ
PQAQLAYQRL	57	IEPQAQLAYQ

H.6 Type 6 Secretion Related Patterns

Table H.31 T6SS-1-15 patterns. A: Highly abundant pattern hits in T6SS-1-15. B. The most abundant pattern and its pairs in the T6SS-1-15.

A-)		B
Pairs	Counts	GYRLLQEYFAAPQRF pattern pairs
GYRLLQEYFAAPQRF	8	GYRLLHEYFCYPEGY
RLLQEYFAAPQRFLF	8	GYRVLQEYLSFPEAF
GYRLLQEYFSLPEKF	8	GYRLLQEYFAAPQRF
YRLLQEYFSLPEKFM	8	ALRLLTEYFAFPKKF
RLLQEYFSLPEKFMF	8	GYRLLQEYFSLPEKF
ADPDVERLLEGFAFL	8	GYRILQEFFCFPEGF
TDPDVERLLEGFAFL	7	GYRILQEYFCYPDAF
DPDVERLLEGFAFLT	7	GYRLLIEQFLCPEKF
PDVERLLEGFAFLTA	7	RLLQEYFAAPQRFLF pattern pairs
CTPVINLFEHDADPI	7	RLLHEYFCYPEGYLF
YRLLQEYFAAPQRFL	7	RLLQEYFAAPQRFLF
FDPDVERLLEGFAFL	7	RLLTEYFAFPKKFDF
DPDVERLLEGFAFLT	7	RLLQEYFSLPEKFMF
LLQEYFSLPEKFMFF	7	NLLHEYAACPERFYF
KSADPDVERLLEGFA	7	RILQEFFCFPEGFLF
SADPDVERLLEGFAF	7	RILQEYFCYPDAFLF
DPDVERLLEGFAFLS	7	RLLIEQFLCPEKFLF
TFSGYRLLHEYFCYP	6	GYRLLQEYFSLPEKF pattern pairs
FSGYRLLHEYFCYPE	6	GYRLLHEYFCYPEGY
DVERLLEGFAFLTAR	6	GYRVLQEYLSFPEAF
VERLLEGFAFLTARL	6	GYRLLQEYFAAPQRF
ERLLEGFAFLTARLR	6	ALRLLTEYFAFPKKF
RLLEGFAFLTARLRE	6	GYRLLQEYFSLPEKF
LLEGFAFLTARLREK	6	GYRILQEFFCFPEGF
LTARLREKVEDEFPE	6	GYRILQEYFCYPDAF

Table H.32 T6SS-2-15 patterns. A. Highly abundant pattern hits in T6SS-2-15. B. The most abundant pattern and its pairs in the T6SS-2-15.

A

Pairs	Counts
FLDIFSHRLTTQFYR	5
DIFSHRLTTQFYRIW	5
GAEFAERYPKVAGRL	5
EFAERYPKVAGRLGM	5
LLQEYFAFPARFQFI	5
FLGLYGPSSPLPTFY	5
LGLYGPSSPLPTFYT	5
GLYGPSSPLPTFYTE	5
LLQDYFYFPQKFHFI	5
LQDYFYFPQKFHFID	5
EDPHVERLLESFALL	5
DPHVERLLESFALLT	5
EDPHTERLIEAFAFL	5
DPHTERLIEAFAFLA	5
AVQWRLISHLSLNYM	5
QWRLISHLSLNYMSI	5
WRLISHLSLNYMSIV	5

B

FLDIFSHRLTTQFYR pattern pairs
FYDIFNHRLLSLYYR
FLDIFSHRLTTQFYR
FYDLFHHRLISLFYR
FLDIFNHRMMTQFYR
FLDIFSHRMTTLFYE
GAEFAERYPKVAGRL pattern pairs
GAEFAERYPKVAGRL
GKAFARHFPKVARRL
GQEFAAQYPKIASRL
GKAFAEQYPKIARRL
SGEFARRYPKIAGRL
LLQEYFAFPARFQFI pattern pairs
LLQEYFAFPARFQFI
LLQDYFYFPQKFHFI
LLHEYFTFPDKFMFF
LLQNYFFFPSMFHFI
LLSEYFSYPDKFLFI
EDPHTERLIEAFAFL pattern pairs
ADPYVERMMEGFAFL
EDPHIERIIESFALV
EDPHTERLIEAFSFL
EDPHVERLLESFALL
EDPHTERLIEAFAFL
QWRLISHLSLNYMSI pattern pairs
TWRLISHLQMNYLSL
RWRLISQLSLNHMLI
IWRLIAHLSLNHMSL
LWRLVSLLSLNYTLL
QWRLISHLSLNYMSI

Table H.33 T6SS-3-15 patterns. A. Highly abundant pattern hits in T6SS-3-15. B. The most abundant pattern and its pairs in the T6SS-3-15.

A		B
Pairs	Counts	SKIAWTEGTFLRPQH pattern pairs
VIWTEGMFLRPHHFQ	8	GKVIWTEGMFLRPHH
IWTEGMFLRPHHFQQ	8	EKVLWGEGLFLRPQH
WTEGMFLRPHHFQQA	8	NKVAWSEGLFLRPQL
VLWGEGLFLRPQHFQ	8	NRIVWSEGMFLRPQH
VAWSEGLFLRPQLFQ	8	SRVMWSEGMFLLPQH
AWSEGLFLRPQLFQQ	8	NKVIWSEGMFLQPQH
WSEGLFLRPQLFQQQ	8	SKIAWTEGTFLRPQH
NRIVWSEGMFLRPQH	8	KIAWTEGTFLRPQHF pattern pairs
RIVWSEGMFLRPQHF	8	KVIWTEGMFLRPHHF
IVWSEGMFLRPQHFQ	8	KVLWGEGLFLRPQHF
VWSEGMFLRPQHFQQ	8	KVAWSEGLFLRPQLF
WSEGMFLRPQHFQQH	8	RIVWSEGMFLRPQHF
SEGMFLRPQHFQQHD	8	RVMWSEGMFLLPQHF
EGMFLRPQHFQQHDR	8	KVIWSEGMFLQPQHL
GMFLRPQHFQQHDRY	8	KIAWTEGTFLRPQHF
SRVMWSEGMFLLPQH	8	IAWTEGTFLRPQHFQ pattern pairs
RVMWSEGMFLLPQHF	8	VIWTEGMFLRPHHFQ
VMWSEGMFLLPQHFQ	8	VLWGEGLFLRPQHFQ
MWSEGMFLLPQHFQY	8	VAWSEGLFLRPQLFQ
WSEGMFLLPQHFQYQ	8	IVWSEGMFLRPQHFQ
SEGMFLLPQHFQYQD	8	VMWSEGMFLLPQHFQ
NKVIWSEGMFLQPQH	8	VIWSEGMFLQPQHLQ
KVIWSEGMFLQPQHL	8	IAWTEGTFLRPQHFQ
VIWSEGMFLQPQHLQ	8	AWTEGTFLRPQHFQQ pattern pairs
IWSEGMFLQPQHLQQ	8	VWSEGMFIGTQHFQQ
WSEGMFLQPQHLQQH	8	IWTEGMFLRPHHFQQ
SEGMFLQPQHLQQHD	8	LWGEGLFLRPQHFQL
EGMFLQPQHLQQHDR	8	AWSEGLFLRPQLFQQ
AWTEGTFLRPQHFQQ	8	VWSEGMFLRPQHFQQ
WTEGTFLRPQHFQQQ	8	MWSEGMFLLPQHFQY
EGTFLRPQHFQQQER	8	IWSEGMFLQPQHLQQ
GTFLRPQHFQQQERY	8	AWTEGTFLRPQHFQQ

Table H.34 T6SS-4-15 patterns. A: Highly abundant pattern hits in T6SS-4-15. B: The most abundant pattern and its pairs in the T6SS-4-15.

A		B
Pairs	Counts	DPYVERLFEGFAFLM pattern pairs
DPYVERLFEGFAFLM	9	DPHVERLLEGCAFLT
DPYVERLFEGFAFMA	9	ENDIEYLFEHFAFLM
PYVERLFEGFAFMAA	9	DPHTERLIEAFAFLC
YVERLFEGFAFMAAR	9	DPHVERLLEGFAFMA
VERLFEGFAFMAARV	9	DPYVERLFEGFAFLM
LLLEYFAFPQKFLFF	8	DDSVERLFQGFSLMM
VERLLEGFAFMAARV	8	DPYVQRLLQSFAFTA
LLQEYFALPEKFLFF	8	DPYVERLFEGFAFMA
RDPYVERLFEGFAFL	8	DPDVERLLEGFAFLT
VERLFEGFAFLMGRL	8	GYRLLLEYFAFPQKF pattern pairs
ERLFEGFAFLMGRLR	8	GSRLLVEYLHFPEKF
EDPHTERLIEAFAFL	7	GYRLLTEFFALPQKF
QSFPGYRLLLEYFAF	7	GYRLLLEYFAFPQKF
GYRLLLEYFAFPQKF	7	GHRLLQEYFALPEKF
YRLLLEYFAFPQKFL	7	GYQLLLEYFTFRPKF
RLLLEYFAFPQKFLF	7	GYRILQEYFAFPERF
LEYFAFPQKFLFFEL	7	GYRLLQEYFVFPAKF
EYFAFPQKFLFFELT	7	EDPHVERLLEGFAFM pattern pairs
SCEDPHVERLLEGFA	7	EDPHVERLLEGCAFL
CEDPHVERLLEGFAF	7	DDPHVERLIQSFAYS
EDPHVERLLEGFAFM	7	EDPHTERLIEAFAFL
DPHVERLLEGFAFMA	7	EDPHVERLLEGFAFM
PHVERLLEGFAFMAA	7	RDPYVERLFEGFAFL
HVERLLEGFAFMAAR	7	SDPYVERLFEGFAFM
ERLLEGFAFMAARVH	7	TDPDVERLLEGFAFL
GHRLLQEYFALPEKF	7	
ARDPYVERLFEGFAF	7	
ERLFEGFAFMAARVQ	7	

Table H.35 T6SS-4-15 patterns. A. Highly abundant pattern hits in T6SS-5-15. B. The most abundant pattern and its pairs in the T6SS-5-15.

A

Pairs	Counts
VLWGEGMFLRPQHFQ	7
VGQEVIVEFIEGDPD	6
GQEVIVEFIEGDPDH	6
QEVIVEFIEGDPDHP	6
EVIVEFIEGDPDHPL	6
VYWHQGMFLQPQHFQ	6
IGQEVLVDFKNGDPD	5
VCWHEGMQLLPQHFQ	5
WHEGMQLLPQHFQLQ	5
IYWHQGMFMQPQHFQ	5
YWHQGMFMQPQHFQL	5
WHQGMFMQPQHFQLA	5
VVWTEGMYMSPQHFQ	5
WTEGMYMSPQHFQAQ	5
GKGFGMIQIPRIGQE	4
KGFGMIQIPRIGQEV	4
FGMIQIPRIGQEVLV	4
GMIQIPRIGQEVLVD	4
MIQIPRIGQEVLVDF	4
IPRIGQEVLVDFKNG	4
PRIGQEVLVDFKNGD	4
RIGQEVLVDFKNGDP	4
GQEVLVDFKNGDPDL	4
QEVLVDFKNGDPDLP	4
EVLVDFKNGDPDLPI	4
EGTEVAVGFEGGDPD	4
GTEVAVGFEGGDPDR	4
TEVAVGFEGGDPDRP	4
EVAVGFEGGDPDRPF	4

B

VLWGEGMFLRPQHFQ pattern pairs

VCWHEGMQLLPQHFQ
IYWHQGMFMQPQHFQ
VAWSKGVFLSPQHLQ
IHWHEGQFLHPHHFQ
VLWGEGMFLRPQHFQ
VVWTEGMYMSPQHFQ
VYWHQGMFLQPQHFQ

VGQEVIVEFIEGDPD pattern pairs

IGQEVLVDFKNGDPD
EGTEVAVGFEGGDPD
VGQEVIVEFIEGDPD
IGQEVVVSFIDGSPD
QGTEVAIAFEEGDPD
IGSEVLVSFIQGNPD

EVIVEFIEGDPDHPL pattern pairs

EVLVDFKNGDPDLPI
EVAVGFEGGDPDRPF
EVIVEFIEGDPDHPL
EVVVSFIDGSPDKPL
EVAIAFEEGDPDRPY
EVLVSFIQGNPDYPV

IYWHQGMFMQPQHFQ pattern pairs

VCWHEGMQLLPQHFQ
IYWHQGMFMQPQHFQ
VLWGEGMFLRPQHFQ
VVWTEGMYMSPQHFQ
VYWHQGMFLQPQHFQ

Table H.36 T6SS-total-15 patterns. A. Highly abundant pattern hits in whole T6SS. B. The most abundant pattern and its pairs in whole T6SS Data.

A

Pairs	Counts
DPHVERLLEGFALLA	28
EDPHVERLLESFALL	27
DPHVERLLESFALLT	27
EDPHVERLIEAVAFL	27
PDPHVERLLEGFALL	27
PHVERLLEGFALLAA	27
VERLLEGFALLAARL	27
VLWGEGLFLRPQHFQ	26
DPYVERLFEGFAFLM	26
VLWGEGLFLRPQHFQ	26
CEDPHVERLLEGFAF	26
EDPHVERLLEGFAFM	26
DPHVERLLEGFAFMA	26
LLQEYFALPEKFLFF	26
CPDPHVERLLEGFAL	26
RLLQEYFSLPEKFMF	25
IVWSEGMFLRPQHFQ	25
RILQEYFAFPERFLF	25
RDPYVERLFEGFAFL	25
DPHVERLIEAVAFLC	25
EDPHTERLIEAFAFL	25
EDPHTERLIEAFAFL	25
DPHTERLIEAFAFLA	25
VLWGEGMFLRPQHFQ	25

B

DPHVERLLEGFALLA pattern pairs
DPDVERLLEGFAFLT
DPHTERLIEAFSFLS
DPYVERLFEGFAFMA
DPHIERIIESFALVT
DPYVERLFEGFAFLM
DPDIERLLEGVAFLT
DPDVERLLEGFAFLS
DPHVERLLESFALLT
DPHVERLIEAVAFLC
DPHVSRLIEGFSLLT
DPHTERLIEAFAFLC
DPDVERLLEGVAFQT
DPHTERLIEAFAFLA
DPHVERMIQSFALLT
DPHVERMLQSFALLA
DRHVEALLEGRAGAL
DPDVERLLEGFAFLT
DPHVERLIQSFAYSA
DPHVERLIEAVAFLN
DPHVERMLEAFSFLA
DPHVERLLEGFAFMA
DPGIERVFEAFALLI
DPFVERLLEGFCFLT
DPDVERLLEGVAFLT

APPENDIX I

PATTERN CLUSTERING

I.1 Clustal Alignment of Type 1 Secretion System Related Patterns

```
-----INASGGILGR          --IEEINAAGGV---
-----INARGGILGR          --IEEINAAGGV---
-----INAKGGILGR           --IEEINAAGGV---
-----INAKGGVLGR          --AEEINAAGGV---
-----INAKGGVLGR          --VEEINAAGGI---
-----INASGGVLGR          --AEEINAAGGI---
-----INAEGGVLGR          LAVEEVNAAG-----
-----INAAGGILGR          LAVEEVNAAG-----
-----INAAGGVLGR          -AVEEVNAAGG----
-----INAAGGVLGR          -AVEEVNAAGG----
-----INAAGGVLGR          --VEEVNAAGGI---
-----INAAGGLLGR          -AVDEVNAAGG----
-----INAAGGLLGK          --VDEVNAAGGV---
----DINAAGGVLG-          -AVEEINAKGG----
----DINAAGGVLG-          -AVEEINAAGG----
----DLNAAGGVLG-          --LALLAAAA-FA--
-----LNAAGGVLGK          ---ALLAAAAVLA--
-----INAAGGVLGK          -----LALAAALALA
-----INAAGGVLGK          ------ALAAALALA
-----VNAAGGINGR            -----LAAALALALS
-----VNAAGGINGR          ------AAALALALS
-----ANAAGGINGR          -----LAGAAALALA
-----INAAGGINGR          ------AGAAALALA
-----FNAAGGVNGR          ----ILAALAALAL-
-----INAAGGIHGR          ----LLAAAAAVAL-
-----INAAGGIHGR           -----LAAAAAVALA
-----INAAGGINGK           ------AAAAAVALA
-----INAAGGINGK          ---AALAATAAVA--
-----INAAGGVNGK          ----ALAATAAVAA-
-----INAAGGVNGK          ---AAVAAAALLA--
----EINAAGGVLG-          ----AVAAAALLAV-
----EINAAGGVLG-          ---LALAAAALLA--
----EINAAGGVLG-          ---LGLAAALAAA—
```

I.1 continued..

----EINAAGGVLG- -----LAAALAAAAP
----EINAAGGLLG- ------AAALAAAAP
---EINAAGGLLG- -----LAAAMAAAAL
----EVNAAGGLLG- ----MAAAALAVAA-
-----VNAAGGLLGR -----AAAALAVAAV
----EVNAAGGILG- ------AAALTAAAA
----EINAAGGILG- -----AAAALTAAGA
----EINAAGGING- ----AAALALAACG-
----EINAAGGING- ----IAALALAAAG-
----EANAAGGING- -----AALALAAAGA
----EINAAGGIAG- ------ALALAAAGA
----EANAAGGVAG- ----AIAAAAAFAA-
----EINAAGGVNG- -----IAAAAAFAAA
----EINAAGGVNG- ----AAALALGLAA-
-AIDEINAAGG---- -----AVLAAGAAAA
-AIDEINAAGG---- ---AALAAAAGLL--
-AIDEINAAGG---- ----ALAAAAGLLA-
-AIDEANAAGG---- -----LALAALAAAG
-ALDEINAAGG---- --GGVLGRKIEL---
-ALDEINAAGG---- --GELVGFDIDL---
-AVDEINAAGG---- --GELVGFDIDL---
LAIDEINAAG----- --GELVGFDVDL---
LAIDEINAAG---- --GELTGFDIDL---
LALDEINAAG----- --GKLAGFDIDL---
LALDEINAAG----- --GKYVGFDIDL---
LAADEINAAG----- --GKLVGFDVDI---
LAIEEINAAG----- ---KLVGFDVDIA--
LAIEEINAAG----- --GKITGFDVDL---
LAIEEINAAG----- --GKLIGFDVDL---
LAAEEINAAG----- ------GFDVDLANA
LAAEEINAAG----- ------GFDVDLANA
-AAEEINAAGG---- ------GFDVDFAKA
-AAEEINAAGG---- ------GFDVDIAKA
-AAEEINAAGG---- -----TGFDVDLAKA
-AAEAINAAGG---- ------GFDVDLAKA
-AAAEINAAGG---- ------GFDVDLARE
-AADEINAAGG---- ------GFDVDLAKE
--IDEINAAGGL--- -----VGFDIDLAKA
--IDEINAAGGL--- ------GFDIDLAKA
--IEEINAAGGL--- ------GFDIDLATA
-AIEEINAAGG---- ------GFDIDLAQL
-AIEEINAAGG---- ------GFDIDIANA
-AIEEINAAGG---- ------GFDIDVARA
-AIEEINAAGG---- ----KIGVLLPLSG-
--FDEINAAGGV--- ----KIGVVLPLSG-
--LDEINAAGGV--- ----KIGVIAPLSG-
--VDEINAAGGV--- ----KIGLLAPLSG-
--IDEINAAGGV--- ----KIGLLAPLTG-
--VDEINAAGGI--- ----KIGVYLPLTG-
--VDEINAAGGI--- ----KIGVIYPLTG-
--LDEINAAGGI--- ----KIGVLTPLTG-
--ADEINAAGGI--- ----KIGALVPLTG-
--IDEINAAGGI--- -----IGVVAPLTGP

I.2 Clustal Alignment of Type 2 Secretion System Related Patterns

```
---VIISGGT--GSGKTTLL------         ---LLVTGPT--GSGKSTTL------
 ---IIISGGT--GSGKTTLL------    ---ILVTGPT--GSGKSTTL------
----IISGGT--GSGKTTLLN-----    ---ILVTGPT--GSGKSTTL------
 ----IISGGT--GSGKTTLLN-----   ---ILVTGPT--GSGKSTTL------
 ---IIISGGT--GSGKTTML------    ---ILVTGPT--GSGKSISL------
 ----IISGGT--GSGKTTMLN-----   ---IIFTGPT--GSGKSTSL------
 ----IISGGT--GSGKTTTLN-----   ---IIVTGPT--GSGKSTTL------
 ---FIISGGT--GSGKTTTL------    ------TGPT--GSGKSTTLVAM---
 ---AIIAGGT--GSGKTTTL------    -------GPT--GSGKSTTLVAMI--
 ----IIAGGT--GSGKTTTLN-----   -----VTGPT--GSGKSTTLVA----
 -----ISGGT--GSGKTTTLNC----    ----LVTGPT--GSGKSTTLV-----
 -----ISGGT--GSGKTTTLNI----   ---YVVAGTT--GSGKSTTL------
 ----LISGGT--GSGKTTTLN-----    ----VVAGTT--GSGKSTTLK-----
 ----LIVGGT--GSGKTTTLN-----    ---IVLSGPT--GSGKSTTL------
 ----LIAGGT--GSGKTTTLN-----    ---IVLSGPT--GSGKSTTL------
 ------AGGT--GSGKTTTLNTI---    ----VLSGPT--GSGKSTTLR-----
 -------GGT--GSGKTTTLNTIA--    ----VLSGPT--GSGKSTTLR-----
 -----IAGGT--GSGKTTTLNT----    -----LSGPT--GSGKSTTLRT----
 -----IVGGT--GSGKTTTLNV----    -----LSGPT--GSGKSTTLRS----
 -----IAGGT--GSGKTTTLNV----    ------SGPT--GSGKSTTLRSA---
 ------VGGT--GSGKTTTLNVV---   ------SGPT--GSGKSTTLRTA---
 -------GGT--GSGKTTTLNVVS--   ----LVTGPT--GSGKSTTLY-----
 --NTLIVGGT--GSGKTTT-------     ----LVTGPT--GSGKSTTLY-----
 --NTLIAGGT--GSGKTTT-------     ----LVTGPT--GSGKSTTLY-----
 ---TLIVGGT--GSGKTTTL------    ----LVTGPT--GSGKSTTLY-----
 ---TLIAGGT--GSGKTTTL------    ----LVTGPT--GSGKSTTLY-----
 --NVLISGGT--GSGKTTT-------    ----LVTGPT--GSGKSTTLY-----
 ---VLISGGT--GSGKTTTL------    ---ILITGAT--GSGKSSTL------
 --NVIVSGGT--GSGKTTT-------    ---VLITGPT--GSGKSSTL------
 ---VIVSGGT--GSGKTTTL------    ----LITGPT--GSGKSSTLY-----
 -ANILVAGGT--GSGKTT--------    ---ILITGPT--GSGKSTTL------
 --NILVAGGT--GSGKTTT-------    ----LITGPT--GSGKSTTLY-----
 ---ILVAGGT--GSGKTTTL------    ---LLFTGPT--GCGKSTTL------
 --NIFVSGGT--GSGKTTT-------    ----LFTGPT--GCGKSTTLY-----
 ---IFVSGGT--GSGKTTTL------    -----VTGPT--GSGKSTTLYA----
------SGGT--GSGKTTLLNCL---    -----VTGPT--GSGKSTTLYA----
 ------SGGT--GSGKTTLLNIL---    -----VTGPT--GSGKSTTLYA----
 ------SGGT--GSGKTTLLNAL---   -----VTGPT--GSGKSTTLYA----
 -----ISGGT--GSGKTTMLNA----    ------TGPT--GSGKSTTLYAG---
 ------SGGT--GSGKTTMLNAL---   ------TGPT--GSGKSTTLYAT---
 ------SGGT--GSGKTTTLNAL---   -----VTGPT--GSGKSTTLYG----
 ------SGGT--GSGKTTTLNCL---     ----IVTGPT--GSGKSTTLY-----
 -------GGT--GSGKTTTLNCLG--     ------TGPT--GSGKSTTLYAA---
 -------GGT--GSGKTTTLNVLS--    ------TGPT--GSGKSTTLYAA---
 -------GGT--GSGKTTTLNVLS--     ------TGPT--GSGKSTTLYGA---
 ------AGGT--GSGKTTTLNVL---    ------TGPT--GSGKSTTLYSA---
```

I.2 continued..

-------GGT--GSGKTTLLNALS-- -----VTGPT--GSGKSTTLYS----
-------GGT--GSGKTTTLNALS-- -----VTGPT--GSGKTTTLYS----
--------GT--GSGKTTTLNALSS- -----VTGPT--GSGKTTTLYS----
---------T--GSGKTTTLNALSSF -----VTGPT--GSGKTTTLYS----
------SGGT--GSGKTTTLNIF--- -----VTGPT--GSGKTTTLYS----
-------GGT--GSGKTTTLNIFS-- ------TGPT--GSGKTTTLYSV---
----FVSGGT--GSGKTTTLN----- ------TGPT--GSGKTTTLYSV---
-----VSGGT--GSGKTTTLNV---- ------TGPT--GSGKTTTLYSS---
----IVSGGT--GSGKTTTLN----- ------TGPT--GSGKTTTLYSL---
-----VSGGT--GSGKTTTLNA---- ------TGPT--GSGKTTTLYSL---
------AGGT--GSGKTTTLNSL--- ------TGPT--GSGKTTTLYST---
-------GGT--GSGKTTTLNSLA-- -----ATGPT--GSGKTTTLYS----
-----VAGGT--GSGKTTTLNS---- -GMLIFTGPT--GSGKTT--------
----LVAGGT--GSGKTTTLN----- --MLIFTGPT--GSGKTTT-------
---LMVAGGT--GSGKTTSL------ ---LIFTGPT--GSGKTTTL------
----MVAGGT--GSGKTTSLN----- ----IFTGPT--GSGKTTTLY-----
------AGGT--GSGKTTSLNAI--- -----FTGPT--GSGKTTTLYS----
-------GGT--GSGKTTSLNAIS-- ------TGPT--GSGKTTTLHSM---
-----VAGGT--GSGKTTSLNA---- -------GPT--GSGKTTTLHSMI--
----LFAGAT--GSGKTTSMN----- -----VTGPT--GSGKTTTLHS----
----IVAGET--ASGKTTTLN----- -----VTGPT--GSGKTTTLYA----
----IVAGGT--ASGKTTTLN----- -----VTGPT--GSGKTTTLYA----
---MIVAGGT--ASGKTTTL------ -----VSGPT--GSGKTTTLYA----
---MLVAGGT--ASGKTTAL------ ---VLVSGPT--GSGKTTTL------
----LVAGGT--ASGKTTALN----- ----LVSGPT--GSGKTTTLY-----
---LLVAGGT--AAGKTTTL------ -GLLLFSGPT--GSGKST--------
----LVAGGT--AAGKTTTLN----- ----LFSGPT--GSGKSTLMY-----
------AGGT--ASGKTTTLNAL--- GVNILVSGPT--GSGKT---------
-------GGT--ASGKTTTLNALS-- -VNILVSGPT--GSGKTT--------
-----VAGGT--ASGKTTTLNA---- --NILVSGPT--GSGKTTL-------
-------GET--ASGKTTTLNAIL-- ---ILVSGPT--GSGKTTLL------
----FVVGET--ASGKTTTLN----- -GLILLSGPM--GSGKTT--------
----LVTGGT--ASGKTSMLN----- --LILLSGPM--GSGKTTT-------
----LVTGPT--GSGKTTTLY---- HGMLLISGPT--GSGKT---------
----LVTGPT--GSGKTTTLY----- -GMLLISGPT--GSGKTT--------
----LVTGPT--GSGKTTTLY----- --MLLISGPT--GSGKTTT-------
----LVTGPT--GSGKTTTLY----- ---LLISGPT--GSGKTTTL------
----LVTGPT--GSGKTTTLY----- ----LISGPT--GSGKTTTLY-----
----LVTGPT--GSGKTTTLY----- -----ISGGT--GSGKTTLLNC----
----LVTGPT--GSGKTTTLY----- -----ISGGT--GSGKTTLLNV----
-------GPT--GSGKSTTLYAAL-- -----ISGGT--GSGKTTLLNA----
-------GPT--GSGKSTTLYAAL-- ----LISGGT--GSGKTTLLN-----
-------GPT--GSGKSTTLYGAL-- ---MLISGGT--GSGKTTLL------
-------GPT--GSGKSTTLYSAL-- ---ILISGGT--GTGKTTLL------
-------GPT--GSGKSTTLYAGL-- ------SGGT--GSGKTTLLNVA---
-------GPT--GSGKSTTLYATL-- -------GGT--GSGKTTLLNVAA--
-------GPT--GSGKSTTLYSIL-- -LNILVSGGT--GSGKTT--------
-------GPT--GSGKSTTLYSML-- --NILVSGGT--GSGKTTL-------
-------GPT--GSGKTTTLYSVL-- -KNLLVSGGT--GSGKTT--------
-------GPT--GSGKTTTLYSVL-- --NLLVSGGT--GSGKTTL-------
-------GPT--GSGKTTTLYSSL-- ----LVSGGT--GSGKTTLMN-----

I.2 continued..

-------GPT--GSGKTTTLYSTL-- -----VSGGT--GSGKTTLMNA---
--------PT--GSGKTTTLYSLVQ- ---ILVSGGT--GSGKTTLM------
--------PT--GSGKTTTLYSLLQ- ----LVSGGT--GSGKTTLLN-----
-------GPT--GSGKTTTLYSLV-- -----VSGGT--GSGKTTLLNI----
-------GPT--GSGKTTTLYSLL-- ---LLVSGGT--GSGKTTLL------
-------GPT--GSGKTTTLYGAL-- ----LVSGPT--GSGKTTLLN-----
-------GPT--GSGKTTTLYTAL-- -----VSGPT--GSGKTTLLNA----
-------GPT--GSGKTTTLYAAL-- ------SGPT--GSGKTTLLNAL---
-------GPT--GSGKTTTLYAAL-- -------GPT--GSGKTTLLNALG--
--------PT--GSGKTTTLYAALN- -------GPT--GAGKTTLLNSLM--
---------T--GSGKTTTLYAALNA ---FLVTGGT--GAGKTTLL------
--------PT--GSGKTTTLYAALH- ----LVTGGT--GAGKTTLLS-----
------SGPT--GSGKTTTLYAS--- ---VLVTGGT--GAGKTTLL------
-----ISGPT--GSGKTTTLYA---- ----LVTGGT--GAGKTTLLK-----
------SGPT--GSGKTTTLYAT--- -----VTGGT--GAGKTTLLKA---
-------GPT--GSGKTTTLYATL-- ------TGGT--GAGKTTLLSAL---
------TGPT--GSGKTSTLYAS--- -------GGT--GAGKTTLLSALL--
-------GPT--GSGKTSTLYASL-- -----VTGGT--GAGKTTLLSA----
-------GPT--GSGKTTTLYASL-- ---VLITGGT--GAGKTTLL------
------TGPT--GSGKTTTLYAM--- ----LITGGT--GAGKTTLLR-----
------TGPT--GSGKTTTLYAM--- -----VSGGT--GAGKTTLLNA----
-------GPT--GSGKTTTLYAML-- -----VSGGT--GAGKTTLLNA----
--------PT--GSGKTTTLYAMLK- ----VVSGGT--GAGKTTLLN-----
-------GPT--GSGKTTTLYAMI-- ----LVSGGT--GAGKTTLLN-----
--------PT--GSGKTTTLYAMIS- ---ILVSGGT--GAGKTTLL------
-GLIVICGPT--GSGKTT-------- -KIILVSGPT--GAGKTT--------
--LIVICGPT--GSGKTTT------- --IILVSGPT--GAGKTTL-------
---IVICGPT--GSGKTTTL------ ---ILVSGPT--GAGKTTLL------
----VICGPT--GSGKTTTLY----- ----LVSGPT--GAGKTTLLL-----
-----ICGPT--GSGKTTTLYA---- -----VSGPT--GAGKTTLLLW----
------CGPT--GSGKTTTLYAA--- ---VLFAGPT--GVGKTTLL------
------TGPT--GSGKTTTLYGA--- ----LFAGPT--GVGKTTLLN-----
------TGPT--GSGKTTTLYTA--- ---VLFSGPT--GAGKTTLL------
------TGPT--GSGKTTTLYAA--- ----LFSGPT--GAGKTTLLN-----
-GIIYITGPT--GSGKTT-------- ---YLIVGST--GSGKTTFL------
--IIYITGPT--GSGKTTT------- ----LIVGST--GSGKTTFLN-----
---IYITGPT--GSGKTTTL------ ---VLVVGST--GSGKSTSL------
----YITGPT--GSGKTTTLY----- ----LRQDPDIIMIGEMRD-------
------TGPT--GSGKTTTLYMI--- ----LRQDPDIIMIGEVRD-------
-------GPT--GSGKTTTLYMIL-- ----LRQDPDIIMIGEIRD-------
-----ITGPT--GSGKTTTLYM---- ----LRQDPDIIMIGEIRD-------
----LLTGPT--GSGKTTTLY----- ---MLRQDPDIIMIGEIR--------
-----LTGPT--GSGKTTTLYA---- ---LLRQDPDIIMIGEIR--------
----LATGPT--GSGKTTTLY----- ---ILRQDPDIIMIGEVR--------
--IILVTGPT--GSGKSTT------- ---ILRHDPDIIMIGEIR--------
--IILVTGPT--GSGKSTT------- ----LRHDPDIIMIGEIRD-------
--IILVTGPT--GSGKSTT------- ---FLRQDPDIIMVGEIR--------
--IILITGPT--GSGKSTT------- ---FLRQDPDIIMVGEIR--------
--IIIVTGPT--GSGKSTT------- ---LLRQDPDIIMVGEIR--------
--IILVTGPT--GSGKTTT------- ----LRQDPDIIMVGEIRD-------

I.2 continued..

```
 --IILATGPT--GSGKTTT-------      ----LRQDPDIIMVGEIRD-------
 --IIFVTGPT--GSGKTTT-------      ----LRQDPNIIMVGEIRD-------
 --ILLVTGPT--GSGKTTT-------      ---FLRQDPDIIMLGEIR--------
 --ILLVTGPT--GSGKTTT-------      ----LRQDPDIIMLGEIRD-------
 NGIVLVSGPT--GSGKT---------      ----LRQDPDIILIGEMRD-------
 -GIVLVSGPT--GSGKTT--------      ---LLRQDPDIILVGETR--------
 --IVLVSGPT--GSGKTTT-------      ----LRQDPDIILVGETRD-------
 --IVLVTGPT--GSGKTTT-------      ---LLRQDPDVLMVGEIR--------
 --IVLLTGPT--GSGKTTT-------      ----LRQDPDVLMVGEIRD-------
 --MILVTGPT--GSGKTTT-------      ----LRQDPDVVMVGEIRD-------
 --MILVTGPT--GSGKTTT-------      ----LRQDPDVVMVGEIRD-------
 --IILVTGPT--GSGKTST-------      ----LRQDPDIVMVGEIRD-------
 ---ILVTGPT--GSGKTSTL------      ----LRQDPDIVMVGEIRD-------
 ---ILVTGPT--GSGKTTTL------      ---ALRRDPDVLMVGEIR--------
 ---ILVTGPT--GSGKTTTL------      ---ALRHDPDILMVGEIR--------
 ---ILVTGPT--GSGKTTTL------      ---LLRQDPDILVVGEIR--------
 ---ILVTGPT--GSGKTTTL------      ----LRQDPDILVVGEIRD-------
 ---ILATGPT--GSGKTTTL------      ---ILRHDPDMIILGEIR--------
 ---IFVTGPT--GSGKTTTL------      ----LRHDPDMIILGEIRD-------
 ---LLVTGPT--GSGKTTTL------      ---ILRHDPDKILVGEIR--------
 ---LLVTGPT--GSGKTTTL------      ----LRHDPDKILVGEIRD-------
 ----LVTGPT--GSGKTTTLH-----      ---ILRHDPDTILVGEIR--------
 SGFFLVTGPT--GSGKT---------      ----LRHDPDTILVGEIRD-------
 -GFFLVTGPT--GSGKTT--------      ----LRHRPDLLIIGEIRD-------
 --FFLVTGPT--GSGKTTT-------      ----LRHDPDILIIGEIRD-------
 ---FLVTGPT--GSGKTTTL------      ---ILRCDPDVILIGEIR--------
 ---VLLTGPT--GSGKTTTL------      ----LRCDPDVILIGEIRD-------
 ---VLVTGPT--GSGKTTTL------      ---ILRQDPDVILIGEIR--------
 ---ILVTGPT--GSGKTVSL------      ----LRQDPDVILIGEIRN-------
 ---ILVTGPT--GSGKTVSL------      ---ILRQDPDVIMIGEIR--------
 ---ILMTGPT--GSGKTVSL------      ---ILRQDPDVIMIGEIR--------
 ----LMTGPT--GSGKTVSLY-----      ----LRQDPDVIMIGEIRD-------
 ----LVTGPT--GSGKTVSLY-----      ----LRQDPDVIMIGEIRD-------
 ----LVTGPT--GSGKTVSLY-----      ----LRQDPDVIMVGEIRD-------
 ---VLVTGPT--GSGKTLSL------      ----LRQDPDVIMVGEIRD-------
 ----LVTGPT--GSGKTLSLY-----      ----LRQDPDVIMVGEIRD-------
 --LLLVTGPT--GSGKTVT-------      ----LRQDPDVIMVGEIRD-------
 ---LLVTGPT--GSGKTVTL------      ---FLRQDPDVIMVGEIR--------
 ----LVTGPT--GSGKTVTLY-----      ---FLRQDPDVIMVGEIR--------
 -----VTGPT--GSGKTVTLYS----      ---FLRQDPDVIMVGEIR--------
 ----LVTGPT--GSGKTSTLY-----      ---ILRQDPDVIMVGEIR--------
 -----VTGPT--GSGKTSTLYA----      ---FLRQDPDVISVGEIR--------
 -----VTGPT--GSGKTTTLYG----      ----LRQDPDVISVGEIRD-------
 -----VTGPT--GSGKTTTLYT----      ---ALRQDPDVILLGELR--------
 ----FVTGPT--GSGKTTTLY-----      ----LRQDPDVILLGELRD-------
-GVVLVTGPT--GSGKST--------      ---ILRQDPDVILLGEIR--------
 -GLVLVTGPT--GSGKST--------      ----LRQDPDVILLGEIRD-------
 -GIVLVTGPT--GSGKST--------      ----LRQAPDVILIGEIRS------
```

I.2 continued..

-GIILITGPT--GSGKST-------- ----LRYRPDMIVVGEIRG-------
EGILLVTGPT--GSGKT--------- ---ALRQRPDYIVMGEIR--------
HGILLVTGPT--GSGKT--------- ---ALRQRPDYIVMGEIR--------
-GILLVTGPT--GSGKTT-------- ---ALRQRPDYIVMGEIR--------
-GILLVTGPT--GSGKTT-------- ---ALRQRPDYIVMGEIR--------
-GIVLLTGPT--GSGKTT-------- ----LRQRPDYIVMGEIRG-------
-GIVLVTGPT--GSGKTT-------- ----LRQRPDYIVMGEIRG-------
HGLVLVTGPT--GSGKT--------- ----LRQRPDYIVMGEIRG-------
HGIVLVTGPT--GSGKT--------- ----LRQRPDYIVMGEIRG-------
-GLVLVTGPT--GSGKTL-------- ---ALRSRPDYIVVGEVR--------
--LVLVTGPT--GSGKTLS------- ----LRSRPDYIVVGEVRG-------
QGLLLVTGPT--GSGKT--------- --NALRQRPDIMLVGEI---------
-GLLLVTGPT--GSGKTV-------- ---ALRQRPDIMLVGEIR--------
QGLILVTGPT--GSGKS--------- ----LRQRPDIMLVGEIRT-------
YGMILVTGPT--GSGKT--------- ---SLRQRPEYILVGEIR--------
YGMILVTGPT--GSGKT--------- ----LRQRPEYILVGEIRT-------
QGMILVTGPT--GSGKT--------- ---ALRQRPNFILVGEIR--------
HGMILVTGPT--GSGKT--------- ----LRQRPNFILVGEIRD-------
-GMILVTGPT--GSGKTV-------- ---ALRQRPNYILVGEIR--------
-GMILVTGPT--GSGKTV-------- ----LRQRPNYILVGEIRG-------
-GMILVTGPT--GSGKTT-------- ---TLRMRPDRIIVGEVR--------
-GMILVTGPT--GSGKTT-------- ---ALRMRPDRIIVGEVR--------
NGIILVTGPT--GSGKT--------- ----LRMRPDRIIVGEVRS-------
QGIILVTGPT--GSGKT--------- ---ALRMRPDRIVVGEVR--------
HGIILVTGPT--GSGKT--------- ---ALRMRPDRILVGEVR--------
-GIILVTGPT--GSGKTS-------- ---ALRMRPDRIVVGECR--------
-GIILVTGPT--GSGKTT-------- ----LRMRPDRIVVGEVRG-------
-GIILVTGPT--GSGKTT-------- ----LRMRPDRILVGEVRG-------
-GIILATGPT--GSGKTT-------- ---ALRMRPDRIILGEIR--------
-GIIFVTGPT--GSGKTT-------- ---ALRMRPDRIILGEIR--------
-GMLLVTGPT--GSGKST-------- ---ALRMRPDRIIIGEVR--------
--MLLVTGPT--GSGKSTT------- ---ALRMRPDRIIIGEIR--------
--VVLVTGPT--GSGKSTT------- ----LRMRPDRIILGEIRG-------
--LVLVTGPT--GSGKSTT------- ----LRMRPDRIILGEIRG-------
--IVLVTGPT--GSGKSTT------- ----LRMRPDRIIIGEIRG-------
---VLVTGPT--GSGKSTTL------ ---ALRMRPDRIIVGETR--------
---VLVTGPT--GSGKSTTL------ ---SLRQRPDRVIVGEVR--------
---VLVTGPT--GSGKSTTL------ ----LRQRPDRVIVGEVRG-------
-GLILVGGAT--GSGKST-------- ----LRMSPDRVIVGEIRG-------
--LILVGGAT--GSGKSTT------- ---SLRMRPDRLVVGEIR--------
-GLLLVTGAT--GSGKST-------- ----LRMRPDRLVVGEIRK-------
--LLLVTGAT--GSGKSTT------- LRAALRQRPDIIIVG-----------
----LVGGAT--GSGKSTTIY----- -RAALRQRPDIIIVGE----------
---LLVTGAT--GSGKSTTL------ --AALRQRPDIIIVGET---------
-----VTGAT--GSGKSTTLAA---- ---ALRQRPDIIIVGETR--------
------TGAT--GSGKSTTLAAM--- ----LRQRPDIIIVGETRG-------
----LVTGAT--GSGKSTTLA----- ---ALRYHPDIICVGEMR--------

I.3 Clustal Alignment of Type 3 Secretion System Related Patterns

```
 LGDIPVVGHL-------      -TKVPFLGDIP------
 LGDIPVLGHL-------      -AKVPFLGDIP------
LGDIPVLGRL-------      --KVPFLGDIPY-----
LGDIPYLGAL-------      --KVPFLGDIPF-----
LGDIPYLGRL-------       --RVPLLADIPL-----
LGDIPILGAL-------       ---VPLLADIPLV----
LGDIPILGEL-------       ----PLLADIPLVG---
LGDIPILGRL-------       ----PLLGDIPLAG---
LGDIPGLGFL-------       -----LLGDIPLAGR--
LGDIPGLGRL-------       ---IPLLGDIPLA----
 -GDIPVVGHLF------      ------LGSIPYLGRL-
 --DIPVVGHLFR-----      -------GSIPYLGRLF
-GDIPVLGHLF------       -----LLGSIPYLGR--
--DIPVLGHLFK-----       ----PLLGSIPYLG---
---IPVLGHLFKS----       -----LLGDIPYLGA--
-GDIPYLGALF------       ----PLLGDIPYLG---
-GDIPYLGRLF------       ----PLLGDIPGLG---
-GDIPGLGFLF------       -----LLGDIPGLGF--
-GDIPGLGRLF------       ----PLLGDIPILG---
-GDIPVLGRLF------       ----PLLGDIPILG---
--DIPYLGALFR-----       -----LLGDIPILGR--
---IPYLGALFRS----       -----LLGDIPILGE--
--DIPYLGRLFR-----       -----LLGDLPLLGA--
---IPYLGRLFRK----       ------LGDLPLLGAL-
--DIPVLGRLFR-----       ------LGQLPLLGRL-
---IPVLGRLFRR----       -------GQLPLLGRLF
--DIPGLGRLFR-----       ----PGLGQLPLLG---
---LPLLGALFRR----       ----PLLGDLPLLG---
---LPLVGALFRS----       ---ALAVRAAAAA----
--DLPLLGALFR-----       ----LAVRAAAAAW---
-GDLPLLGALF------       ---AAARRAAAAA----
--ELPVLGALFR-----       ----AARRAAAAAL---
---LPVLGALFRS----       ---GAAEAAAAAA----
-GELPVLGALF------       ----AAEAAAAAAG---
 --DLPVLGRLFK-----      ---AGAAAAAAAV----
 ---LPVLGRLFKS----      ----GAAAAAAAVP---
 -SDLPVLGRLF------      --AAGAAAAAAA-----
---LPLLGRLFSS----       -----AAAAAAAVPL--
 -GDVPALGHLF------       ----LAFAAADAAA---
 --DVPALGHLFR-----      ----AGAGAPAAAP---
LGDVPALGHL-------            ----AAPAAAVPAP---
 -GDIPILGALF------      ---PAAVPAAVPA----
 --DIPILGALFS-----       ----AAVAASAAPA---
--DIPILGRLFK-----        ----AALAAVAAPL---
--DIPILGELFK-----        ----AAVAAAIARV---
-GDIPILGELF------        --AAAATAATAA-----
-GDIPILGRLF------        --AAAAGAASAA-----
 -GDVPILGPLF------       --AAFACAAIAA-----
 --DVPILGPLFR-----       ----AAIAAIAAGL---
--DMPILGNLFR-----        -----AIAAIAAGLA--
---MPILGNLFRS----       --ACAAIAAIAA-----
```

I.3 continued..

-GDMPILGNLF------	--AAAALAACVA-----
LGDMPILGNL-------	---AAALAACVAA----
-RDIPFLGRLF------	----AALAACRAAA---
--DIPFLGRLFD-----	-----ALAACRAAAA--
LRDIPFLGRL-------	---AAALMAPAAA----
-GDIPFLGNLF------	---VAARAAALMA----
--DIPFLGNLFK-----	----AARAAALMAP---
LGDIPFLGNL-------	---RAARLAALVA----
LADIPLVGAL-------	---RAVRGAALAA----
-ADIPLVGALF------	----AVRGAALAAA---
--DIPLVGALFK-----	----AALAATLLAP---
---IPLVGALFKR----	---AAARARLARA----
-GDIPLAGRLF------	----AIRAALARAL---
--DIPLAGRLFQ-----	---AAAPAALADA----
LGDIPLAGRL-------	----RARAAAALAA---
LGSIPLIGGL-------	-----ARAAAALAAC--
-GSIPLIGGLF------	---ALARADAALA----
LGSIPFIGKL-------	--RAAAAGWLAA-----
-GSIPFIGKLF------	---AAAAGWLAAR----
-GDIPFIGSLF------	----AAAGWLAARL---
--DIPFIGSLFR-----	---AAAAGALAAR----
LGDIPFIGSL-------	---AAAAGALAER----
--DIPVIGALFR-----	--CAAAAGALAA-----
---IPVIGALFRS----	-----AVSALAERLA--
-GDIPVIGALF------	-----AVSALAKRLA--
LGDIPVIGAL-------	-----AAAALAKRLA--
-----LLGDIPVVGH--	----AAASALAERV---
-----LLGDIPVLGH--	-----AASALAERVV--
-----LLGDIPVLGR--	-----AASALAERLV--
----PLLGDIPVLG---	-----AADALAGRLA--
----PLLGDIPVLG---	-----AAGALAGRLA--
----PLLGDIPVVG---	-----AAGALAGRLV--
----PGLGSIPLIG---	----AAAGALAARL---
----PILGSIPFIG---	----AAAGALADRL---
----PFLSSIPVIG---	----AAAGALAERL---
----PLLGDIPVIG---	-----AAGALAERLA--
-----LLGDIPVIGA--	-----AAGALAARLA--
----PLLGDIPFIG---	------AGALAARLAV-
-----LLGDIPFIGS--	----AALSALAARL---
-NKIPLLGDIP------	--AAALALSLAA-----
--KIPLLGDIPL-----	--AAAPAGALAA-----
-RKIPLLGDIP------	--ALAVGGALAA-----
NDKIPLLGDI-------	--LLVAGVALAA-----
NSKIPLLGDI-------	---LVAGVALAAL----
-DKIPLLGDIP------	----VAGVALAALA---
-SKIPLLGDIP------	---PVA-LAAAAAG---
--KIPLLGDIPV-----	----VA-LAAAAAGG--
---IPLLGDIPVV----	-----A-LAAAAAGGL-
---IPLLRDIPFL----	-------LAAAAAGGLA
----PLLRDIPFLG---	-AIPVA-LAAAA-----
---IPLLGDIPFI----	---LLALAAGPAA----

I.3 continued..

--KIPLLGDIPF-----	---ALALGLAVAA----
--KIPLLGSIPY-----	----LALGLAVAAA---
---IPLLGSIPYL----	-----ALGLAVAAAL--
--KIPILGSIPF-----	----LALGAAAAVL---
-AKIPYLGDIP------	---APAPAAGAAA----
--KIPYLGDIPI-----	----PAPAAGAAAA---
----PYLGDIPILG---	---APAAGAAAAA----
-----YLGDIPILGA--	----PAAGAAAAAA---
---IPYLGDIPIL----	---APLVAAFAAA----
--RTPILGDIPI-----	---IGLLLAILLA----
---TPILGDIPIV----	----GLLLAILLAL---
----PILGDIPIVQ---	-----LLLAILLALP--
---IPLLGDIPIL----	---LAGLLALGLA----
---FPLLGDIPIL----	---ALVLLALRLL----
--KIPLLGDIPI-----	---AAVLLAVMLA----
ESKVPLLGDI-------	---AIDLLAARLA----
VEKVPLLGDI-------	----GVLLAALLAG---
ADKVPLLGDI-------	-----VLLAALLAGG--
-DKVPLLGDIP------	-AALFVLLAAA------
-EKVPLLGDIP------	--ALFVLLAAAL-----
-SKVPLLGDIP------	---LFVLLAAALA----
-RKVPLLGDIP------	---LVVLLVALLL----
--KVPLLGDIPG-----	--FSLALFLALL-----
--KVPLLGDIPY-----	---SLALFLALLA----
---VPLLGDIPVL----	----LALFLALLAL---
---VPLLGDIPVL----	-----ALFLALLALV--
---VPLLGDIPGL----	---ALVLALALLA----
---VPLLGDIPYL----	----LVLALALLAG---
-DSVPLLGDIP------	--AALVLALALL-----
--SVPLLGDIPV-----	--VLALLLLALL-----
---VPLLGDIPVI----	---LALLLLALLT----
--KVPLLGDIPV-----	-GVLALLLLAL------
--KVPLLGDIPV-----	---LLLLLLPLAV----
-NKFPLLGDIP------	----LLLLLPLAVA---
--KFPLLGDIPI-----	----LLLLLVLVLL---
-SKVPFLGDVP------	----VALLAALDAL---
--KVPFLGDVPA-----	--AAAALFGALL-----
---VPFLGDVPAL----	----LLLIALVLAG---
----PFLGDVPALG---	----LALGALLLAG---
-DKVPFLGDLP------	------LGALLVAGVA-
--KVPFLGDLPI-----	---SLALNAFSKW----
---VPFLGDLPII----	---AIALNALSKW----
----PFLGDLPIIG---	--LSNALNALSK-----
----PFLGDLPVIG---	--LSQALNALSK-----
--KLPFLGDLPV-----	---SNVLNALSKW----
---LPFLGDLPVI----	------LNALSKWPDT-
-SKLPFLGDLP------	------LNALSKWPDT-
-DKVPMLGDMP------	------LNALSKWPDT-
--KVPMLGDMPI-----	-------NALSKWPDTQ
---VPMLGDMPIL----	------LNALSKWPRT-
----PMLGDMPILG---	------LNALSKWPRT-

I.3 continued..

VDKVPLLGDL-------	------LNALSKWADN-
-DKVPLLGDLP------	----QALNALSKWP---
-NKVPLLGDLP------	----QVLNALSKWP---
--KVPLLGDLPI-----	-----ALDAVSKWPD--
--KVPLLGDLPL-----	-----VLSALSRWPD--
---VPLLGDLPLL----	--VTQALNALSK-----
---VPLLGDLPII----	--VTQVLNALSK-----
----PLLGDLPIIG---	--VTSALNALSK-----
-QSIPFLGDIP------	-GVANLLNALS------
--SIPFLGDIPG-----	--VANLLNALSK-----
---IPFLGDIPGL----	-DVANVLNALS------
-----FLGDIPGLGR--	--VANVLNALSK-----
-----FLGDIPYLGR--	-GVAVALNALS------
----PFLGDIPGLG---	--VAVALNALSK-----
----PFLGDIPYLG---	-GVANALNALS------
----PFLGDIPFLG---	-GVVVVLNALS------
-----FLGDIPFLGN--	--VVVVLNALSK-----
---VPFLGDIPYL----	----GQTVVLGGLI---
---VPFLGDIPFL----	---DGETVVIGGV----
QTKVPFLGDI-------	---KTELVIFVTP----

I.4 Clustal Alignment of Type 4 Secretion System Related Patterns

-----TVPGQVSCRVSQDVY----------
------VPGQVSCRVSQDVYS----------
-------PGQVSCRVSQDVYSA----------
--------GQVSCRVSQDVYSAD----------
---------QVSCRVSQDVYSADG----------
----------VSCRVSQDVYSADGL----------
-----------SCRVSQDVYSADGLV---------
------------CRVSQDVYSADGLVR--------
-------------RVSQDVYSADGLVRL-------
--------------VSQDVYSADGLVRLI------
-----DYDGFVTCRVTQDVY----------
------YDGFVTCRVTQDVYS----------
-------DGFVTCRVTQDVYSS----------
--------GFVTCRVTQDVYSSN----------
---------FVTCRVTQDVYSSNG----------
----------VTCRVTQDVYSSNGA----------
-----------TCRVTQDVYSSNGAV---------
------------CRVTQDVYSSNGAVL--------
--------TELDTTVPGQVSCRV------------
--------SPNGQHIVRHVADPF------------
---------PNGQHIVRHVADPFS-----------
-------TSPNGQHIVRHVADP-------------
----------NGQHIVRHVADPFSL----------
-----------GQHIVRHVADPFSLA---------
------QGDAVSIFVARDLDF--------------
-------GDAVSIFVARDLDFS-------------
---------AVSIFVARDLDFSGV-----------
----------VSIFVARDLDFSGVY----------
-----------SIFVARDLDFSGVYT---------
------------IFVARDLDFSGVYTL--------
-------------FVARDLDFSGVYTLA-------
--------------VARDLDFSGVYTLAD------
-----------NIMVVRDVDFSTVYR---------
------------IMVVRDVDFSTVYRL--------
----------INIMVVRDVDFSTVY----------
-------------MVVRDVDFSTVYRLE-------
--------------VVRDVDFSTVYRLEG------
-RVTQDVYSSNGAVLL--------------------
--VTQDVYSSNGAVLLV-------------------
---TQDVYSSNGAVLLVE------------------
----QDVYSSNGAVLLVER-----------------
-----DVYSSNGAVLLVERG----------------
------VYSSNGAVLLVERGS--------------
-------YSSNGAVLLVERGSL-------------
--------SSNGAVLLVERGSLV------------
---------SNGAVLLVERGSLVS-----------
----------NGAVLLVERGSLVSG----------
------------AVLLVERGSLVSGTQ--------
-------------VLLVERGSLVSGTQK-------
-----------GAVLLVERGSLVSGT---------
RSQGKIVFVDEAWQL--------------------

------------PMILTKKARDDLKKL--------
-------------MILTKKARDDLKKLK-------
--------------ILTKKARDDLKKLKL------
---------------LTKKARDDLKKLKLI-----
--GARDFILSHGSSIPC------------------
---ARDFILSHGSSIPCA-----------------
----RDFILSHGSSIPCAL----------------
-----DFILSHGSSIPCALY---------------
------FILSHGSSIPCALYT--------------
-------ILSHGSSIPCALYTQ-------------
--------LSHGSSIPCALYTQI------------
----------HGSSIPCALYTQIIS----------
-------ASQIDTSLLTKLSSL-------------
--------SQIDTSLLTKLSSLK------------
IGDFFENSLQYPCPY--------------------
-GDFFENSLQYPCPYI-------------------
-------SLQYPCPYIISMGIH-------------
--------LQYPCPYIISMGIHI------------
------NSLQYPCPYIISMGI--------------
---------QYPCPYIISMGIHIL-----------
----------YPCPYIISMGIHILD----------
--TYHLHKGGGMCELYH------------------
---YHLHKGGGMCELYHT-----------------
----HLHKGGGMCELYHTI----------------
-----LHKGGGMCELYHTIG---------------
------HKGGGMCELYHTIGI--------------
-------KGGGMCELYHTIGIF-------------
---------GGMCELYHTIGIFAP-----------
----------GMCELYHTIGIFAPR----------
--------GGGMCELYHTIGIFA------------
-----------MCELYHTIGIFAPRS---------
------------CELYHTIGIFAPRSQ--------
-------------ELYHTIGIFAPRSQL-------
--------------LYHTIGIFAPRSQLD------
--------VVAIIGVLAAVAIPA------------
---------VAIIGVLAAVAIPAY-----------
----------AIIGVLAAVAIPAYQ----------
-----------IIGVLAAVAIPAYQN---------
------------IGVLAAVAIPAYQNY--------
-------------GVLAAVAIPAYQNYV-------
---------------LAAVAIPAYQNYVQK-----
----------------AAVAIPAYQNYVQKT----
--------------VLAAVAIPAYQNYVQ------
--------VIAIVGILAAVALPA------------
---------IAIVGILAAVALPAY-----------
----------AIVGILAAVALPAYQ----------
-----------IVGILAAVALPAYQD---------
------------VGILAAVALPAYQDY--------
-------------GILAAVALPAYQDYT-------
----------------AAVALPAYQDYTARA----
-----------------AVALPAYQDYTARAQ---

I.4 continued..

-SQGKIVFVDEAWQLL---------- ----------LAAVALPAYQDYTAR-----
--QGKIVFVDEAWQLLD---------- ----------ILAAVALPAYQDYTA------
---GKIVFVDEAWQLLDD---------- --------VVAIIGILAAFAIPA----------
----KIVFVDEAWQLLDDT---------- ---------VAIIGILAAFAIPAY----------
-----IVFVDEAWQLLDDTE---------- ----------AIIGILAAFAIPAYN----------
------VFVDEAWQLLDDTEE---------- ----------IIGILAAFAIPAYND---------
-------FVDEAWQLLDDTEET---------- ----------IGILAAFAIPAYNDY--------
--------VDEAWQLLDDTEETA---------- ----------GILAAFAIPAYNDYI-------
---------DEAWQLLDDTEETAA---------- ----------AAFAIPAYNDYIART----
----------EAWQLLDDTEETAAF---------- ----------AFAIPAYNDYIARTQ---
-----------AWQLLDDTEETAAFI--------- ----------LAAFAIPAYNDYIAR-----
------------WQLLDDTEETAAFIE-------- ----------ILAAFAIPAYNDYIA------
-------------QLLDDTEETAAFIEE------- ----------AIIGILAAIAIPQYQ----------
--------------LLDDTEETAAFIEEG------ ----------IIGILAAIAIPQYQD---------
------QEAMDGRRFVLDIDE---------- ----------IGILAAIAIPQYQDY--------
-------EAMDGRRFVLDIDEA---------- ----------GILAAIAIPQYQDYT-------
--------AMDGRRFVLDIDEAW---------- ----------ILAAIAIPQYQDYTA------
---------MDGRRFVLDIDEAWK---------- ----------LAAIAIPQYQDYTAR-----
----------DGRRFVLDIDEAWKY---------- ----------AIAIPQYQDYTARTQ---
-----------GRRFVLDIDEAWKYL--------- ----------IAIPQYQDYTARTQV--
------------RRFVLDIDEAWKYLG-------- ----------AAIAIPQYQDYTART----
-------------RFVLDIDEAWKYLGD------- ---------IVVLGILAVTALPRL----------
--------------FVLDIDEAWKYLGDP------ ----------VVLGILAVTALPRLL----------
---------DEAWKYLGDPKVAYF---------- --------VIVVLGILAVTALPR----------
----------DKVPLLGDIPVIKRL---------- ----------VLGILAVTALPRLLN---------
-----------KVPLLGDIPVIKRLF--------- ----------LGILAVTALPRLLNL--------
------------VPLLGDIPVIKRLFS-------- ----------GLASVENLIALSSAT----------
---------QVDTHMWERLR-GAIM---------- ----------LASVENLIALSSATL---------
----------VDTHMWERLR-GAIMI--------- --------IPQYQDYTARTQVTR----------
-----------DTHMWERLR-GAIMIS-------- ---------PQYQDYTARTQVTRA----------
---------WVDNHYFERFS-GAIM---------- -------AIPQYQDYTARTQVT----------
----------VDNHYFERFS-GAIML--------- ----------QYQDYTARTQVTRAV----------
-----------DNHYFERFS-GAIMLS-------- ----------YQDYTARTQVTRAVS---------
---------VRDMLKTARKR-NAIV---------- ----------QDYTARTQVTRAVSE--------
----------RDMLKTARKR-NAIVR--------- ----------DYTARTQVTRAVSEV-------
------------MLKTARKR-NAIVRLA------- ------VALPAYQDYTARAQV----------
-------------LKTARKR-NAIVRLAT------ -------ALPAYQDYTARAQVS----------
--------------KTARKR-NAIVRLATQ----- --------LPAYQDYTARAQVSE----------
---------------TARKR-NAIVRLATQS---- ---------PAYQDYTARAQVSEA----------
-----------------RKR-NAIVRLATQSIT-- ----------AYQDYTARAQVSEAI----------
------------------KR-NAIVRLATQSITD- ----------YQDYTARAQVSEAIL---------
---------------ARKR-NAIVRLATQSI--- ----------QDYTARAQVSEAILL--------
------ICFYLFARIQEAMD-G---------- ----------DYTARAQVSEAILLA-------
-------CFYLFARIQEAMD-GR---------- ------FAIPAYNDYIARTQV----------
--------FYLFARIQEAMD-GRR---------- -------AIPAYNDYIARTQVS----------
---------YLFARIQEAMD-GRRF---------- --------IPAYNDYIARTQVSE----------
----------LFARIQEAMD-GRRFV--------- ---------PAYNDYIARTQVSEG----------
-----------FARIQEAMD-GRRFVL-------- ----------AYNDYIARTQVSEGV----------
------------RIQEAMD-GRRFVLDI------ ----------YNDYIARTQVSEGVS---------
-------------IQEAMD-GRRFVLDID----- ----------NDYIARTQVSEGVSL--------
------------ARIQEAMD-GRRFVLD------- ----------DYIARTQVSEGVSLA-------

I.4 continued..

---------IMLEAMGGGFTWGAI----------- --------------YIARTQVSEGVSLAD------
----------MLEAMGGGFTWGAIL---------- ---------------IARTQVSEGVSLADG-----
-----------LEAMGGGFTWGAILI--------- ----------------ARTQVSEGVSLADGL----
------------EAMGGGFTWGAILIR-------- -----------------RTQVSEGVSLADGLK---
-------EEDNRIAFRG-FGVMR------------ --------------YTARAQVSEAILLAE------
--------EDNRIAFRG-FGVMRY----------- ---------------TARAQVSEAILLAEG-----
---------DNRIAFRG-FGVMRYP---------- ----------------ARAQVSEAILLAEGQ----
----------NRIAFRG-FGVMRYPK--------- -----------------RAQVSEAILLAEGQK---
-----DLEEDNRIAFRG-FGV-------------- ------------TNKGMSMTEATYALM--------
------------IAFRG-FGVMRYPKFK------- ------------TNKTLSKTEADYNLL--------
-------------AFRG-FGVMRYPKFKH------ ----------EKICMNPFTWLEVIE----------
-----------RIAFRG-FGVMRYPKF-------- -----------KICMNPFTWLEVIED---------
--------------FRG-FGVMRYPKFKHL----- -----MNNEMIYLELEELKD---------------
---------------RG-FGVMRYPKFKHLY---- ------NNEMIYLELEELKDS--------------
----------------G-FGVMRYPKFKHLYE--- -------NEMIYLELEELKDSP-------------
------------------FGVMRYPKFKHLYEM-- --------EMIYLELEELKDSPE------------
-------------------GVMRYPKFKHLYEMG- ---------MIYLELEELKDSPEL-----------
#AD? ----------IYLELEELKDSPELK----------
---------AYRVEFR--YPADEARA--------- -----------YLELEELKDSPELKT---------
----------YRVEFR--YPADEARAK-------- ------------LELEELKDSPELKTV--------
----LDDTEETAAFIEEGY---------------- -------------ELEELKDSPELKTVV-------
-----DDTEETAAFIEEGYR--------------- --------------LEELKDSPELKTVVL------
------DTEETAAFIEEGYRR-------------- ---------------EELKDSPELKTVVLF-----
-------TEETAAFIEEGYRRA------------- ----------------ELKDSPELKTVVLFG----
--------EETAAFIEEGYRRAR------------ -----------------LKDSPELKTVVLFGV---
---------ETAAFIEEGYRRARK----------- ------------------KDSPELKTVVLFGVM--
----------TAAFIEEGYRRARKY---------- -----------DLELSALERENNVEI---------
-----------AAFIEEGYRRARKYF--------- ------------LELSALERENNVEII--------
------------AFIEEGYRRARKYFG-------- ----------LDLELSALERENNVE----------
-------------FIEEGYRRARKYFGS------- ----EFREAVLGLEVTPHI----------------
---IEEGYRRARKYFGS-F---------------- -----FREAVLGLEVTPHIS---------------
----EEGYRRARKYFGS-FG--------------- ------REAVLGLEVTPHISK--------------
-----EGYRRARKYFGS-FGM-------------- -------EAVLGLEVTPHISKD-------------
------GYRRARKYFGS-FGMG------------- --------AVLGLEVTPHISKDN------------
-------YRRARKYFGS-FGMGT------------ ---------VLGLEVTPHISKDNN-----------
--------RRARKYFGS-FGMGTQ----------- ----------LGLEVTPHISKDNNI----------
---------RARKYFGS-FGMGTQG---------- -----------GLEVTPHISKDNNIL---------
----------ARKYFGS-FGMGTQGI--------- ------------LEVTPHISKDNNILL--------
------------KYFGS-FGMGTQGIDD------- -------------EVTPHISKDNNILLD-------
-------------YFGS-FGMGTQGIDDA------ -----HQKRELVIFVTPHIL---------------
-----------RKYFGS-FGMGTQGID-------- ------QKRELVIFVTPHILK--------------
----FGSFGMGTQGIDDAF---------------- -------KRELVIFVTPHILKA-------------
-----GSFGMGTQGIDDAFA--------------- --KQKKQQGFTLIELMI------------------
------SFGMGTQGIDDAFAN-------------- ---QKKQQGFTLIELMIV-----------------
-------FGMGTQGIDDAFAND------------- ----KKQQGFTLIELMIVV----------------
--------GMGTQGIDDAFANDA------------ -----KQQGFTLIELMIVVA---------------
---------MGTQGIDDAFANDAA----------- ------QKGFTLIELMIVVAI--------------
----------GTQGIDDAFANDAAR---------- ------QKGFTLIELMIVVAI--------------
-----------TQGIDDAFANDAARA--------- ------QKGFTLIELMIVVAI--------------
------------QGIDDAFANDAARAA-------- ------QQGFTLIELMIVVAI--------------
-------------GIDDAFANDAARAAY------- -------KGFTLIELMIVVAII-------------

I.4 continued..

----------IDDAFANDAARAAYN------ -------KGFTLIELMIVVAII----------
----------DDAFANDAARAAYNS----- -------KGFTLIELMIVVAII----------
----------DAFANDAARAAYNSS---- -------QGFTLIELMIVVAII----------
----------AFANDAARAAYNSSD--- --------GFTLIELMIVVAIIG----------
--FANDAARAAYNSSDW---------- --------GFTLIELMIVVAIIG----------
---ANDAARAAYNSSDWK---------- --------GFTLIELMIVVAIIG----------
----NDAARAAYNSSDWKF---------- --------GFTLIELMIVVAIIG----------
-----DAARAAYNSSDWKFF---------- ---------FTLIELMIVVAIIGI----------
-------ARAAYNSSDWKFFLR---------- ---------FTLIELMIVVAIIGI----------
--------RAAYNSSDWKFFLRQ---------- ---------FTLIELMIVVAIIGI----------
------AARAAYNSSDWKFFL---------- ---------FTLIELMIVVAIIGV----------
---------AAYNSSDWKFFLRQD---------- ------QKGFTLIELMIVIAI----------
----------AYNSSDWKFFLRQDE---------- -------KGFTLIELMIVIAIV----------
-----------YNSSDWKFFLRQDEQ--------- ---MKAQKGFTLIELMIV----------
------------NSSDWKFFLRQDEQS-------- ----KAQKGFTLIELMIVV----------
-------------SSDWKFFLRQDEQSF------- -----AQKGFTLIELMIVVA----------
--------------SDWKFFLRQDEQSFE------ -----AQKGFTLIELMIVVA----------
---------------DWKFFLRQDEQSFEK----- ----AAQKGFTLIELMIVV----------
----------------WKFFLRQDEQSFEKL---- --MKAAQKGFTLIELMI----------
----RIKQARSTQNAESKM---------- ---KAAQKGFTLIELMIV----------
-----IKQARSTQNAESKMA---------- --MKSLQKGFTLIELMI----------
------KQARSTQNAESKMAK---------- ---KSLQKGFTLIELMIV----------
-------QARSTQNAESKMAKW---------- ----SLQKGFTLIELMIVV----------
--------ARSTQNAESKMAKWQ---------- -----LQKGFTLIELMIVVA----------
---------RSTQNAESKMAKWQP---------- ----TLQKGFTLIELMIVI----------
----------STQNAESKMAKWQPE---------- -----LQKGFTLIELMIVIA----------
-----------TQNAESKMAKWQPEY--------- --MNTLQKGFTLIELMI----------
------------QNAESKMAKWQPEYA-------- ---NTLQKGFTLIELMIV----------
-------------NAESKMAKWQPEYAE------- ---MKNKQGFTLIELVIV----------
--------------AESKMAKWQPEYAEI------ ----KNKQGFTLIELVIVI----------
---------------ESKMAKWQPEYAEIA----- -----NKQGFTLIELVIVIV----------
----------------SKMAKWQPEYAEIAR---- ------KQGFTLIELVIVIVV----------
-----------------KMAKWQPEYAEIARD--- -------QGFTLIELVIVIVVL----------
------------------MAKWQPEYAEIARDW-- --------GFTLIELVIVIVVLG----------
---LKLITTKTTINAVDM---------- ----------VIVIVVLGILAVTAL-----
----KLITTKTTINAVDMM---------- ----------IVIVVLGILAVTALP----
-----LITTKTTINAVDMMP---------- ----------LVIVIVVLGILAVTA------
------ITTKTTINAVDMMPI---------- ----------ELVIVIVVLGILAVT-------
-------TTKTTINAVDMMPIL---------- ----------IELVIVIVVLGILAV--------
--------TKTTINAVDMMPILS---------- ----------LIELVIVIVVLGILA---------
----------TTINAVDMMPILS-EW--------- ----------TLIELVIVIVVLGIL----------
-----------TINAVDMMPILS-EWQ-------- ---------FTLIELVIVIVVLGI----------
---------KTTINAVDMMPILS-E---------- -MNAKPCRGFTLIELI----------
------------INAVDMMPILS-EWQG------- --NAKPCRGFTLIELIM----------
-------------NAVDMMPILS-EWQGF------ ---AKPCRGFTLIELIMV----------
--------------AVDMMPILS-EWQGFG----- ----KPCRGFTLIELIMVI----------
---------------VDMMPILS-EWQGFGK---- -----PCRGFTLIELIMVIV----------
----------------DMMPILS-EWQGFGKP--- ------CRGFTLIELIMVIVL----------
------------IVNIDSAGTNSLGSA-------- -------RGFTLIELIMVIVLL----------
-------------VNIDSAGTNSLGSAG------- ----------TLIELIMVIVLLAVV----------
----------GTIVNIDSAGTNSLG---------- -----------LIELIMVIVLLAVVS---------

I.4 continued..

----------IVRINSLGTGQLGAA--------
----------VRINSLGTGQLGAAG--------
----------GVIVRINSLGTGQLG----------
-----NKQGNYNFYICGTSG----------
------KQGNYNFYICGTSGA----------
-------QGNYNFYICGTSGAG----------
--------GNYNFYICGTSGAGK----------
---------NYNFYICGTSGAGKS----------
----------YNFYICGTSGAGKSV----------
-----------NFYICGTSGAGKSVF---------
----------FYICGTSGAGKSVFS--------
----------YICGTSGAGKSVFSL-------
----------GNLGLGGGAGSGGGG----------
-----------NLGLGGGAGSGGGGS---------
--------SAGHTLILGSTGSGK----------
---------AGHTLILGSTGSGKT----------
-------ASAGHTLILGSTGSG----------
----------GHTLILGSTGSGKTV----------
-----------HTLILGSTGSGKTVF---------
----------TLILGSTGSGKTVFM--------
----------LILGSTGSGKTVFMS-------
---------KSNILLVGPTGCGKT----------
----------SNILLVGPTGCGKTY----------
-----------NILLVGPTGCGKTYL---------
----------ILLVGPTGCGKTYLA--------
----------LLVGPTGCGKTYLAQ-------
----------LSVTRSGSYKD--PQLI--------
-----------SVTRSGSYKD--PQLIK-------
---------ILSVTRSGSYKD--PQL---------
-----------TNGMYGSYFNGKANI---------
----------NGMYGSYFNGKANID--------
----------GMYGSYFNGKANIDM-------
--------------YGSYFNGKANIDMNN-----
----------ASEALRSYMSIPPTL----------
-----------SEALRSYMSIPPTLY---------
---------LASEALRSYMSIPPT----------
----------IPPTLYDQQGDAVSI----------
----LPLLYSSMPMILTKK----------
-----PLLYSSMPMILTKKA----------
---QLPLLYSSMPMILTK----------
------LLYSSMPMILTKKAR----------
-------LYSSMPMILTKKARD----------
--------YSSMPMILTKKARDD----------
---------SSMPMILTKKARDDL----------
----------SMPMILTKKARDDLK----------
-----------MPMILTKKARDDLKK---------

---------FTLIELIMVIVLLAV----------
--------GFTLIELIMVIVLLA----------
----------TLIELMIVVAIIGIL----------
----------TLIELMIVVAIIGIL----------
----------TLIELMIVVAIIGIL----------
----------TLIELMIVVAIIGVL----------
-----------LIELMIVVAIIGILA---------
-----------LIELMIVVAIIGILA---------
-----------LIELMIVVAIIGVLA---------
---------------MIVVAIIGVLAAVAI-----
---------------IVVAIIGVLAAVAIP----
-------------LMIVVAIIGVLAAVA------
------------ELMIVVAIIGVLAAV-------
-----------IELMIVVAIIGVLAA--------
-----------IELMIVVAIIGILAA--------
-----------IELMIVVAIIGILAA--------
------------ELMIVVAIIGILAAI-------
------------ELMIVVAIIGILAAF-------
-------------LMIVVAIIGILAAIA------
-------------LMIVVAIIGILAAFA------
---------------MIVVAIIGILAAIAI-----
---------------MIVVAIIGILAAFAI-----
---------------IVVAIIGILAAIAIP----
---------------IVVAIIGILAAFAIP----
----------------VVAIIGILAAIAIPQ---
-----------------VAIIGILAAIAIPQY--
--------GFTLIELMIVIAIVG----------
---------FTLIELMIVIAIVGI----------
---------------MIVIAIVGILAAVAL-----
---------------IVIAIVGILAAVALP----
-------------LMIVIAIVGILAAVA------
------------ELMIVIAIVGILAAV-------
-----------IELMIVIAIVGILAA--------
----------LIELMIVIAIVGILA---------
----------TLIELMIVIAIVGIL----------
---------MNKIGLLIVAGVLGL----------
----------NKIGLLIVAGVLGLA----------
-----------KIGLLIVAGVLGLAG---------
-----------IGLLIVAGVLGLAGC--------
-------MDKPIEQIAIEARIV----------
--------DKPIEQIAIEARIVT----------
-------IEARIVTITDESLKE----------
---------LGSAGIPGQVDTHMW----------
--------VDKYFSFSKALAVSQ----------
------------ETLGKIIAKNEKILL--------
------------TQVFAKDGETIVLGG--------

I.5 Clustal Alignment of Type 5 Secretion System Related Patterns

--------KGAVDGYSVGIYGTW------ ----------DASATDKLVVKGDTA
---------GAVDGYSVGIYGTWF----- ----------LDDDASATDKLVVKG---
----------AVDGYSVGIYGTWFA---- ---------TVLDDDASATDKLVV-----
----------VDGYSVGIYGTWFAD---- -----LTLDTVLGDDDSATD---------
--------RGSVDGYNLGIYATW------- ------TLDTVLGDDDSATDR--------
---------GSVDGYNLGIYATWF------ -------LDTVLGDDDSATDRL-------
---------GKVNGYSIGGYATWY------ --------DTVLGDDDSATDRLV------
---------GSVEGYSIGGYATWY------ ----------LGDDDSATDRLVING---
----------SVNGYSTGLYATWYA----- ----------GDDDSATDRLVINGD--
----------VNGYSTGLYATWYAD---- ----------VLGDDDSATDRLVIN----
-------SRGSVDGYSVGLYAT-------- ---------TVLGDDDSATDRLVI-----
--------RGSVDGYSVGLYATW------ ----------LGDDDSATDKLVITG---
---------GSVDGYSVGLYATWY----- ----------GDDDSATDKLVITGD--
----------SVDGYSVGLYATWYA---- ----------VLGDDDSATDKLVIT----
----------VDGYSVGLYATWYAN--- ---------TVLGDDDSATDKLVI-----
-----------DGYSVGLYATWYANE-- -----LYLNTVLGDDDSATD---------
-------SKGSVRGYSAGLYAT------- ------YLNTVLGDDDSATDK--------
--------KGSVRGYSAGLYATW------ -------LNTVLGDDDSATDKL-------
---------GSVRGYSAGLYATWF----- --------NTVLGDDDSATDKLV------
----------SVRGYSAGLYATWFA---- -----LIINTVLGDDTSTTD---------
----------VRGYSAGLYATWFAD--- ------IINTVLGDDTSTTDK--------
-----------RGYSAGLYATWFADD-- -------INTVLGDDTSTTDKL-------
-----------GYSAGLYATWFADDI- ----------LGDDTSTTDKLIVTG---
--------EGTVSGYSAGLYATW------ ----------GDDTSTTDKLIVTGN--
---------GTVSGYSAGLYATWF----- ----------VLGDDTSTTDKLIVT----
----------TVSGYSAGLYATWFQ---- ---------TVLGDDTSTTDKLIV-----
----------VSGYSAGLYATWFQN--- --------NTVLGDDTSTTDKLI------
---------GQVTGYSVGLYGTWY----- -----ATDKLVITGDASGTT---------
----------QVTGYSVGLYGTWYA---- ------TDKLVITGDASGTTD--------
--------RGQVTGYSVGLYGTW------- -----ATDRLVINGDATGTT---------
--------DGSVAGYSVGLYGTW------ -----ATDKVVINGNTSGTT---------
---------GSVAGYSVGLYGTWL----- -------DKVVINGNTSGTTRV-------
----------VTGYSVGLYGTWYAN--- ----SVTDKLVVEGDTSGT----------
-----------TGYSVGLYGTWYANN-- -----VTDKLVVEGDTSGTT---------
-----------NGYSVGLYGTWYANQ-- ------TDKLVVEGDTSGTTA--------
-----------GYSVGLYGTWYANQK- -------DKLVVEGDTSGTTAV-------
----------INGYSVGLYGTWYAN--- -----LTDKLIVKGDTSGNT---------
---------GDINGYSVGLYGTWY----- ---------NLAAANTLFVMRLND-----
----------DINGYSVGLYGTWYA---- ----------LAAANTLFVMRLNDR----
--------RGDINGYSVGLYGTW------ ------------AANTLFVMRLNDRAG--
----------ISGYSVGLYGTWLQD--- ---------NLQAANTLFVHRLHD-----
-----------SGYSVGLYGTWLQDN-- ----------LQAANTLFVHRLHDR----
--------DGRISGYSAGIYATW------ ------------AANTLFVHRLHDRLG--
---------GRISGYSAGIYATWY----- ----------IAAANTLFNTRLHDR----
----------ISGYSAGIYATWYQN--- -----------AAANTLFNTRLHDRL---
--------RGQISGYSAGLYATW------ ---------NIAAANTLFNTRLHD-----
---------GQISGYSAGLYATWY----- ------------AANTLFNTRLHDRLG--
----------ISGYSAGLYGTWYQN--- ------------AANTMFNTRLHDRLG--
-----------SGYSAGLYGTWYQNE-- ---------NLQAANTMFNTRLHD-----
---------GRISGYSAGLYGTWY----- ----------LQAANTMFNTRLHDR----

I.5 continued..

----------ISGYSAGLYATWYGN--- ----------AANTMFTTQLHDRLG--
----------SGYSAGLYATWYGND-- ----------NTMFTTQLHDRLGET
----------GYSAGVYGTWYANDA- ----------AANTLFTMSLHDRLG--
----------GYSVGVYGTWYANDA- ----------ANTLFTMSLHDRLGE-
----------SGYSAGVYGTWYAND--- --------TLTVQGNYVGNNGQL------
---------GEISGYSVGVYGTWY----- ---------LTVEGNYVGNGGTIV-----
----------EISGYSVGVYGTWYA---- --------TLTVAGDYTGNGGHL------
----------ISGYSVGVYGTWYAN--- --------TLTVTGDYAGNGGTL------
----------SGYSVGVYGTWYAND-- ------GNNLTINGDYTGNNG--------
---------GTLSGYSAGVYGTWY----- --------NLTINGDYTGNNGNL------
-------EKYDSKGFTASVEGG-------- -----GSGTLVLGGANTYTG---------
--------KYDSKGFTASVEGGY------- -------GTLVLGGANTYTGGT-------
---------YDSKGFTASVEGGYA----- -----GAGKLTLSGANTYSG---------
--------DYDSRGVTASVEGGY------ -------GKLTLSGANTYSGDT-------
---------YDSRGVTASVEGGYT----- ---------LTLSGANTYSGDTNV-----
-------EKYKSKGITASVEAG------- ---------LTLSGDNTYSGGTTI-----
--------KYKSKGITASVEAGY------- -------GANTYTGGTTVEAGT-------
---------YKSKGITASVEAGYS------ --------ANTYTGGTTVEAGTL------
----------KSKGITASVEAGYSF----- ---------NTYTGGTTVEAGTLI-----
-------ENYKSKGVTASVEAG------- -------GDNSYSGGTTIIGGT-------
--------NYKSKGVTASVEAGY------ -------GDNSYSGGTTIIGGT-------
---------YKSKGVTASVEAGYT----- --------DNSYSGGTTIIGGTL------
----------KSKGVTASVEAGYTW---- --------DNSYSGGTTIIGGTL------
-------EKYKSDGITASVESG------- ---------NSYSGGTTIIGGTLT-----
--------KYKSDGITASVESGY------ ---------NSYSGGTTIIGGTLT-----
---------YKSDGITASVESGYS----- ---------NTYSGGTTITGGTLT-----
---------YKSDGITASVEGGYS----- --------VNTYTGKTTINGGTL------
-------ESYKSDGITASVEGG------- ---------NTYTGKTTINGGTLR-----
--------SYKSDGITASVEGGY------ --------ANTYSGDTNVQEGTL------
-------ENYKSDGFTASVETG------- -SGTTAVTVNNAGGTG----------
--------NYKSDGFTASVETGY------ --GTTAVTVNNAGGTGA----------
---------YKSDGFTASVETGYT----- -SGTTNVTVVNAGGAG----------
----------KSDGFTASVETGYTH----- -TGTTSVRVNNAGGLG----------
-------EKYDSDGITASVETG------- --GTTRVKVTNAGGSGA----------
--------KYDSDGITASVETGY------ --GTTYVTINNLGGQGA----------
---------YDSDGITASVETGYT------ -SGNTFVAVNNIGGAG----------
-------ESYKSKGFTASLEAG-------- --GNTFVAVNNIGGAGA----------
-------ESYKSRGFTASLEAG-------- ----TFVAVNNIGGAGAQT----------
--------SYKSKGFTASLEAGY------- --------VNNIGGAGAQTIEGI------
---------YKSKGFTASLEAGYK------ ---------NNIGGAGAQTIEGIE-----
--------SYKSRGFTASLEAGY------- ----------NIGGAGAQTIEGIEI----
---------YKSRGFTASLEAGYT------ -----------IGGAGAQTIEGIEIV---
----------KSRGFTASLEAGYTQ---- ----------NIGGAGAQTINGMEI----
--------SYRSKGLTASLEAGY------ --------VDNIGGVGAQTVNGI------
---------YRSKGLTASLEAGYT----- ---------DNIGGVGAQTVNGIE-----
-----------SKGLTASLEAGYTLK--- ----------NIGGVGAQTVNGIEL----
---------YKSSGMTASLELGYT----- -----------IGGVGAQTVNGIELI---
-------------GFTASVEGGYAFKVG- ------------GGVGAQTVNGIELIE--
-------------GITASVEAGYSFRLG- -------------GVGAQTVNGIELIEV-
--------AGRINATSTDAINGS------- --------VNNIGGVGARTFEGI------
--------AGRINATSTDAINGS------ ----------NAGGSGAYTLNGIEI----

I.5 continued..

---------GQISATSTDAINGSQ----- ----------AGGSGAYTLNGIEII---
----------QISATSTDAINGSQL---- --------VTNAGGSGAYTLNGI------
--------AGQISATSTDAINGS------ --------VNNAGGTGAKTLNGI------
----------NISLTSTDAINGSQL---- ----TRVSVANAGGGGAQT----------
----------ISLTSTDAINGSQLY--- --------VANAGGGGAQTVEGI------
----------ISATSTDAINGSQLY---- -----GAGAQTIEGIEIVNV---------
----------ISATSTDAINGSQLY---- -------GAQTIEGIEIVNVAG-------
----------ISETSTDAINGSQLY--- -------GAQTVEGIEIVNVGG-------
---------GNISLTSTDAINGSQ----- INNLGGQGAQTVEGI--------------
---------GAISAASSDAINGSQ----- -NNLGGQGAQTVEGIE-------------
------VAAGQISATSTDAIN-------- --NLGGQGAQTVEGIEI------------
-------AAGQISATSTDAING------- ---LGGQGAQTVEGIEIV-----------
-----NVAAGQISATSTDAI--------- ----GGQGAQTVEGIEIVN----------
--------------STDAVNGSQLYAVS -----GQGAQTVEGIEIVNV---------
----------LSEESTDAVNGSQLF--- -------GAYTLNGIEIISVEG-------
-------GRIVAGAYEYKLGRG------- -----GSGQLNKNGTGTLTL---------
-------GRIVAGAYDYYLGRG------- -----GSGQLIKTGQGTLTL---------
-------GRIVAGAYDYTLARG------- -----GS--ITKTGDGTLTLSG-------
-------GRIVAGAYDYTLARG------- ------GQLKTGTNRYVVQLG--------
--------RPEAGSYIANLIAMN------ -------QLKTGTNRYVVQLGG-------
---------PEAGSYIANLIAMNT----- -------QLNTQSNRYVVQLGG-------
-------YRPEAGSYISNIAAA------- ------GQLKTQSNRYVLQLG--------
-------YRPEAGSYIANIAAA------- ------GQLKTQSNRYVLQLG--------
--------RPEAGSYISNIAAAN------ ------GQLKTQANRYVLQLG--------
--------RPEAGSYISNIAAAN------ ------GQLNTQANRYVLQLG--------
---------PEAGSYISNIAAANT----- -------QLKTQSNRYVLQLGG-------
--------RPEAGSYIANIAAAN------ -------QLKTQSNRYVLQLGG-------
---------PEAGSYIANIAAANT----- -------QLKTQANRYVLQLGG-------
--------RPEAGSYTANIAAAN------ ----SSGQLKTQANRYVLQ----------
-------LRPEAGSYIANLAAA------- -----SGQLKTQANRYVLQL---------
--------RPEAGSYIANLAAAN------ ------------SNRYVVQLGGSIAQW--
---------PEAGSYIANLAAANT----- ------------SNRYVTQLGGDVAQW--
----------EAGSYIANLAAANTM---- ------------ANRYVLQIGGDLAQW--
-----------AGSYIANLAAANTMF---- --------GLYVDSWLQYGWYDN------
--------RSEAGSYVANLAAAN------ --------GWYVDSWAQYGWYDN------
----------EAGSYVANLAAANTL----- ----------YVDSWLQYGWYDNTV----
-----------AGSYVANLAAANTLF--- ----------YVDSWAQYGWYDNSV----
------------GSYVANLAAANTLFV--- ---------AYVDSWVQYSWFDNN-----
----------EAGSYIANIAAANTL----- ----------YVDSWVQYSWFDNNV----
-----------AGSYIANIAAANTLF---- --------GAYVDSWVQYSWFDN------
----------EAGSYTANIAAANTL----- -----------VDSWVQYSWFDNNVS---
-----------AGSYTANIAAANTLF---- ------------DSWVQYSWFDNNVSG--
------------GSYIANIAAANTLFN-- ----------YLDSWAQYSWFDNTV----
------------GSYIANLAAANTMFT-- -----------LDSWAQYSWFDNTVK---
-------------SYIANLAAANTMFTT- ------------DSWAQYSWFDNTVKG--
--------------YIANLAAANTMFTTR- ------DTGAYVDSWVQYSWF--------
---------------IANLAAANTMFTTRL -------TGAYVDSWVQYSWFD-------
--------RPEFGSYLANNYAAN------- -----DDTGAYVDSWVQYSW---------
--------RPEFGSYLANNYAAN------ ------KSGLYVDSWVQYNWF--------
--------RPEVGSYLANNYAAN------ -------SGLYVDSWVQYNWFK-------
--------RPEFGSYLANARAAN------ ---------LYVDSWVQYNWFKNR-----

I.5 continued..

--------RPETGSYLANTLVAN------ ----------YVDSWVQYNWFKNRI----
--------GVSLIQVAGKATKDS------ --------GLYVDSWVQYNWFKN------
---------VSLIQVAGKATKDSF------ --------GLYMDAWLQYSWFNN------
------------GGTLLFNTQLGDDSS--- ----------YMDAWLQYSWFNNTV----
------------GGTLIINTVLGDDTS-- ---------AYIDAWAQYSWFKNS-----
------GNYTGNGGSLYLNTV-------- ----------YIDAWAQYSWFKNSV----
-------NYTGNGGSLYLNTVL------- --------GAYIDAWAQYSWFKN------
--------YTGNGGSLYLNTVLG------ ------KKGAYIDAWAQYSWF--------
----------GNGGSLYLNTVLGDD---- ------RNGAYLDSWAQYSWF--------
------------GGSLYLNTVLGDDDS-- -------NGAYLDSWAQYSWFD-------
-------NYVGNGGTIVLNTVL------- --------GAYLDSWAQYSWFDN------
--------YVGNGGTIVLNTVLG------ ---------AYLDSWAQYSWFDNT-----
---------VGNGGTIVLNTVLGG----- ------QEGAYVDTWAQYSWF--------
----------GNGGTIVLNTVLGGD---- -------EGAYVDTWAQYSWFD-------
----------GNNGTIVLNTWLGGD---- --------GAYVDTWAQYSWFDN------
------------GGTIVLNTVLGGDDS-- ---------AYVDTWAQYSWFDNT-----
-----IVLNTVLGGDDSLTD--------- ------KQGVYLDSWAQYGWF--------
------VLNTVLGGDDSLTDK-------- -------QGVYLDSWAQYGWFN-------
-------LNTVLGGDDSLTDKL------- --------GVYLDSWAQYGWFNN------
----------VLGGDDSLTDKLIVK---- ----------YLDSWAQYGWFNNEV----
-----------LGGDDSLTDKLIVKG--- ------------DSWAQYGWFNNEVKG--
---------TVLGGDDSLTDKLIV----- --------GAWLDSWLQYAWFNN------
--------NTVLGGDDSLTDKLI------ -------NGAYVDTWIQYGWFN-------
--------NSELGGDDSLTDKLV------ --------GAYVDTWIQYGWFNN------
-----------LGGDDSLTDKLVIQG--- ---------AYVDTWIQYGWFNNT-----
-----------LGGDNSLTDKLTISG--- ----------YVDTWIQYGWFNNTV----
-----------LGEDNSPTDHLTITG--- -----------VDTWIQYGWFNNTVN---
-----------LGGDNSPTDKMNVKG--- ------------DTWIQYGWFNNTVNG--
-----------LGGDDSPTDKLIVHG--- -------TGGYLDTWMLYSWFN-------
-----LFFNTRLGGDNSLTD--------- --------GGYLDTWMLYSWFNN------
------FFNTRLGGDNSLTDK-------- ----------YLDTWMLYSWFNNAV----
-------FNTRLGGDNSLTDKL------- -----------LDTWMLYSWFNNAVS---
--------NTRLGGDNSLTDKLT------ ------------DTWMLYSWFNNAVSG--
---------TRLGGDNSLTDKLTI----- --------GSYVDTWAAYSWYNN------
-----IVLNTWLGGDNSPTD--------- --------GAYVDTWMLYNWFDN------
------VLNTWLGGDNSPTDK-------- --------GAYVDAWMLYNWFDN------
-------LNTWLGGDNSPTDKV------- --------GAYVDSWMLYNWFDN------
--------NTWLGGDNSPTDKVI------ --------GAYVDSWMLYNWFKN------
-----IAMNTALGGDDSPTD--------- -------TGAYVDAWMLYNWFD-------
------AMNTALGGDDSPTDK-------- -------TGAYVDSWMLYNWFD-------
-------MNTALGGDDSPTDKL------- ------KTGAYVDSWMLYNWF--------
--------NTALGGDDSPTDKLI------ ------KSGAYVDSWMLYNWF--------
---------TALGGDDSPTDKLIV----- --------GAYVDAWALYNWFDN------
-------MNTVLNGDDSPTDKL------- ---------AYVDAWALYNWFDNT-----
-------MNTVLNGDDSVTDKL------- ------KRGAYVDAWALYNWF--------
--------NTVLNGDDSVTDKLV------ -----NKTGAYVDSWALYNW---------
---------TVLNGDDSVTDKLVV----- ------KTGAYVDSWALYNWF--------
---------TVLNGDDSPTDKLLI----- ------KTGWYVDSWALYNWF--------
-----------LNGDDSPTDKLLIKG--- -------TGAYVDSWALYNWFD-------
-----LEISTVLGDDRSPTD--------- --------GAYVDSWALYNWFDN------
------EISTVLGDDRSPTDK-------- --------GWYVDSWALYNWFDN------

I.5 continued..

-------ISTVLGDDRSPTDKL------- ------KQGLYVDSWALYNWF--------
--------STVLGDDRSPTDKLV------ -------QGLYVDSWALYNWFN-------
------------GDDRSPTDKLVVKGD-- --------GLYVDSWALYNWFNN------
-------------DDRSPTDKLVVKGDT- ----------YVDTWMLYNWFDNKV----
-----------LGDDRSPTDKLVVKG--- ----------YVDAWMLYNWFDNQV----
----------VLGDDRSPTDKLVVK---- ----------YVDSWMLYNWFKNTV----
---------TVLGDDRSPTDKLVV----- ---------AYVDSWALYNWFDNS-----
---------TVLGADDSPSDKLVV----- ----------YVDSWALYNWFDNSV----
-----------LGADDSPSDKLVVNG--- ---------AYVDSWMLYNWFDNS-----
-------LQTVLGADDSPSDKL------- ----------YVDSWMLYNWFDNSV----
-------FNTQLGNDDSPTDRM------- ----------YVDAWALYNWFDNTV----
-------FNTQLGSDNSPTDLL------- -----------VDAWALYNWFDNTVN---
-------FNTVLNDDDSETDRL------- ------------DAWALYNWFDNTVNG--
--------NTILAGDTSVTDRLV------ --------GTYLDSWVLYNWFDN------
-----LLFNTQLGDDSSATD--------- ----------YLDSWVLYNWFDNTV----
------LFNTQLGDDSSATDK-------- ------------DSWVLYNWFDNTVRG--
-------FNTQLGDDSSATDKL------- ---------LYVDSWALYNWFNNS-----
--------NTQLGDDSSATDKLI------ ----------YVDSWALYNWFNNSV----
-----------LGDDSSATDKLIIRG--- -----------VDSWALYNWFNNSVT---
------------GDDSSATDKLIIRGD-- ------------DSWALYNWFNNSVTG--
----------QLGDDSSATDKLIIR---- ------------DSWLLYNWFNNSVQG--
---------TQLGDDSSATDKLII----- --------SREFQPFVEANWIHN------
-----MVMNTVLGDDSSLTD--------- ---------REFQPFVEANWIHNT-----
------VMNTVLGDDSSLTDK-------- --------GRDFQPFVEANWIHN------
-------MNTVLGDDSSLTDKL------- --------SREFQPFVEVNWLHN------
--------NTVLGDDSSLTDKLV------ ---------REFQPFVEVNWLHNS-----
--------------DSSLTDKLVVKGNTS ----------EFQPFVEVNWLHNSK----
---------------SSLTDKLVVKGNTS -----------FQPFVEVNWLHNSKD---
------------DDSSLTDKLVVKGNT- --------QRDFQPFVEVNWIHN------
------------GDDSSLTDKLVVKGN-- --------QREFQPFVEVNWIHN------
-----------LGDDSSLTDKLVVKG--- ---------RDFQPFVEVNWIHNS-----
----------VLGDDSSLTDKLVVK---- --------DRTFQPFVEVNWIHN------
---------TVLGDDSSLTDKLVV----- ---------RTFQPFVEVNWIHNT-----
-----IHFNTVLGDDSSLTD--------- ---------REFQPFVEVNWIHNS-----
------HFNTVLGDDSSLTDR-------- ----------EFQPFVEVNWIHNSE----
-------FNTVLGDDSSLTDRM------- -----------FQPFVEVNWIHNSET---
---------TQLGDDSSQTDRMIV----- --------NLKIEPQAQVIYQYL------
-----------LGDDSSQTDRMIVNG--- ---------LKIEPQAQVIYQYLN-----
-----------LGDDSSLTDRMKITG--- ----------KIEPQAQVIYQYLNL----
-------LQTVLGDENSATDKL------- -----------IEPQAQVIYQYLNLE---
--------QTVLGDENSATDKLV------ ------------EPQAQVIYQYLNLED--
---------TVLGDENSATDKLVV----- --------SWRLEPQAQVIYQYL------
--------RTELGDDNSATDKVV------ ---------WRLEPQAQVIYQYLH-----
---------TELGDDNSATDKVVI----- ----------FIQPQAQLTWMGVTA----
-----------LGDDNSATDKVVING--- -----------IQPQAQLTWMGVTAS---
------------GDDNSATDKVVINGN-- ---------YYIQPVAQLTWMGVN-----
-------------DDASATDKLVVKGDT- ----------YIQPVAQLTWMGVNA----

I.6 Clustal Alignment of Type 6 Secretion System Related Patterns

-------ISTVLGDDRSPTDKL------- ------KQGLYVDSWALYNWF--------
--------STVLGDDRSPTDKLV------ -------QGLYVDSWALYNWFN-------
------------GDDRSPTDKLVVKGD-- --------GLYVDSWALYNWFNN------
-------------DDRSPTDKLVVKGDT- ----------YVDTWMLYNWFDNKV----
-----------LGDDRSPTDKLVVKG--- ----------YVDAWMLYNWFDNQV----
----------VLGDDRSPTDKLVVK---- ----------YVDSWMLYNWFKNTV----
---------TVLGDDRSPTDKLVV----- ---------AYVDSWALYNWFDNS-----
---------TVLGADDSPSDKLVV----- ----------YVDSWALYNWFDNSV----
-----------LGADDSPSDKLVVNG--- ---------AYVDSWMLYNWFDNS-----
-------LQTVLGADDSPSDKL------- ----------YVDSWMLYNWFDNSV----
-------FNTQLGNDDSPTDRM------- ----------YVDAWALYNWFDNTV----
-------FNTQLGSDNSPTDLL------- -----------VDAWALYNWFDNTVN---
-------FNTVLNDDDSETDRL------- ------------DAWALYNWFDNTVNG--
--------NTILAGDTSVTDRLV------ --------GTYLDSWVLYNWFDN------
-----LLFNTQLGDDSSATD--------- ----------YLDSWVLYNWFDNTV----
------LFNTQLGDDSSATDK-------- ------------DSWVLYNWFDNTVRG--
-------FNTQLGDDSSATDKL------- ---------LYVDSWALYNWFNNS-----
--------NTQLGDDSSATDKLI------ ----------YVDSWALYNWFNNSV----
-----------LGDDSSATDKLIIRG--- -----------VDSWALYNWFNNSVT---
------------GDDSSATDKLIIRGD-- ------------DSWALYNWFNNSVTG--
----------QLGDDSSATDKLIIR---- ------------DSWLLYNWFNNSVQG--
---------TQLGDDSSATDKLII----- --------SREFQPFVEANWIHN------
-----MVMNTVLGDDSSLTD--------- ---------REFQPFVEANWIHNT-----
------VMNTVLGDDSSLTDK-------- --------GRDFQPFVEANWIHN------
-------MNTVLGDDSSLTDKL------- --------SREFQPFVEVNWLHN------
--------NTVLGDDSSLTDKLV------ ---------REFQPFVEVNWLHNS-----
-------------DSSLTDKLVVKGNTS ----------EFQPFVEVNWLHNSK----
---------------SSLTDKLVVKGNTS -----------FQPFVEVNWLHNSKD---
------------DDSSLTDKLVVKGNT- --------QRDFQPFVEVNWIHN------
------------GDDSSLTDKLVVKGN-- --------QREFQPFVEVNWIHN------
-----------LGDDSSLTDKLVVKG--- ---------RDFQPFVEVNWIHNS-----
----------VLGDDSSLTDKLVVK---- --------DRTFQPFVEVNWIHN------
---------TVLGDDSSLTDKLVV----- ---------RTFQPFVEVNWIHNT-----
-----IHFNTVLGDDSSLTD--------- ---------REFQPFVEVNWIHNS-----
------HFNTVLGDDSSLTDR-------- ----------EFQPFVEVNWIHNSE----
-------FNTVLGDDSSLTDRM------- -----------FQPFVEVNWIHNSET---
---------TQLGDDSSQTDRMIV----- --------NLKIEPQAQVIYQYL------
-----------LGDDSSQTDRMIVNG--- ---------LKIEPQAQVIYQYLN-----
-----------LGDDSSLTDRMKITG--- ----------KIEPQAQVIYQYLNL----
-------LQTVLGDENSATDKL------- -----------IEPQAQVIYQYLNLE---
--------QTVLGDENSATDKLV------ ------------EPQAQVIYQYLNLED--
---------TVLGDENSATDKLVV----- --------SWRLEPQAQVIYQYL------
--------RTELGDDNSATDKVV------ ---------WRLEPQAQVIYQYLH-----
---------TELGDDNSATDKVVI----- ----------FIQPQAQLTWMGVTA----
-----------LGDDNSATDKVVING--- -----------IQPQAQLTWMGVTAS---
------------GDDNSATDKVVINGN-- ---------YYIQPVAQLTWMGVN-----
-------------DDASATDKLVVKGDT- ----------YIQPVAQLTWMGVNA----

APPENDIX K

WEB-LOGO PROJECTION OF CLUSTURED PATTERNS

K.1 Web-Logo Projection of Type 1 Secretion Related Patterns

Figure K.1 The Web-Logo projection of pattern 1 (P1) in TISS

,

Figure K.2 The Web-Logo projection of pattern 2 (P2) in TISS

Figure K.3 The Web-Logo projection of pattern 3 (P3) in TISS

Figure K.4 The Web-Logo projection of pattern 4 (P4) in TISS

Figure K.5 The Web-Logo projection of pattern 5 (P5) in TISS

Figure K.6 The Web-Logo projection of pattern 6 (P6) in TISS

K.2 Web-Logo Projection of Type 2 Secretion Related Patterns

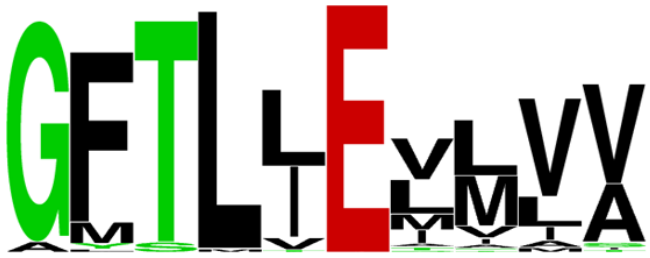


Figure K.7 The Web-Logo projection of pattern 1 (P1) in T2SS

Figure K.8 The Web-Logo projection of pattern 2 (P2) in T2SS

Figure K.9 The Web-Logo projection of pattern 3 (P3) in T2SS

Figure K.10 The Web-Logo projection of pattern 4 (P4) in T2SS

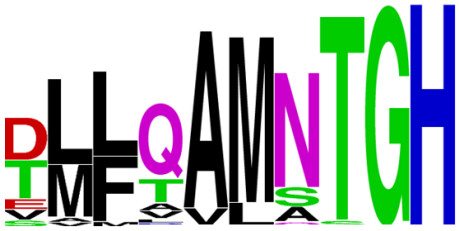


Figure K.11 The Web-Logo projection of pattern 5 (P5) in T2SS

Figure K.12 The Web-Logo projection of pattern 6 (P6) in T2SS

Figure K.13 The Web-Logo projection of pattern 7 (P7) in T2SS

Figure K.14 The Web-Logo projection of pattern 8 (P8) in T2SS

Figure K.15 The Web-Logo projection of pattern 9 (P9) in T2SS

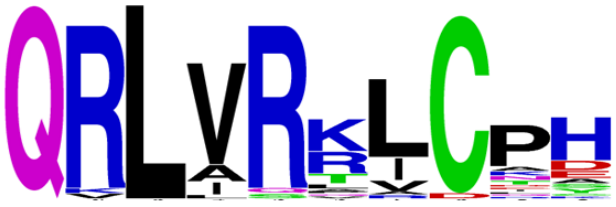


Figure K.16 The Web-Logo projection of pattern 10 (P10) in T2SS

K.3 Web-Logo Projection of Type 3 Secretion Related Patterns

Figure K.17 The Web-Logo projection of pattern 1 (P1) in T3SS

Figure K.18 The Web-Logo projection of pattern 2 (P2) in T3SS

Figure K.19 The Web-Logo projection of pattern 3 (P3) in T3SS

Figure K.20 The Web-Logo projection of pattern 4 (P4) in T3SS

Figure K.21 The Web-Logo projection of pattern 5 (P5) in T3SS

Figure K.22 The Web-Logo projection of pattern 6 (P6) in T3SS

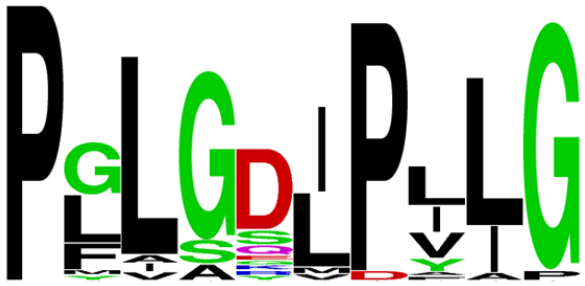


Figure K.23 The Web-Logo projection of pattern 7 (P7) in T3SS

Figure K.24 The Web-Logo projection of pattern 8 (P8) in T3SS

K.4 Web-LogoPprojection of Type 4 Secretion Related Patterns

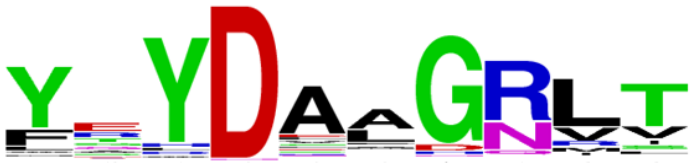


Figure K.25 The Web-Logo projection of pattern 1 (P1) in T4SS

Figure K.26 The Web-Logo projection of pattern 2 (P2) in T4SS

Figure K.27 The Web-Logo projection of pattern 3 (P3) in T4SS

Figure K.28 The Web-Logo projection of pattern 4 (P4) in T4SS

Figure K.29 The Web-Logo projection of pattern 5 (P5) in T4SS

Figure K.30 The Web-Logo projection of pattern 6 (P6) in T4SS

Figure K.31 The Web-Logo projection of pattern 7 (P7) in T4SS

Figure K.32 The Web-Logo projection of pattern 8 (P8) in T4SS

Figure K.33 The Web-Logo projection of pattern 9 (P9) in T4SS

Figure K.34 The Web-Logo projection of pattern 10 (P10) in T4SS

Figure K.35 The Web-Logo projection of pattern 11 (P11) in T4SS

Figure K.36 The Web-Logo projection of pattern 12 (P12) in T4SS

Appendix K.5 Web-Logo Projection of Type 5 Secretion Related Patterns

Figure K.37 The Web-Logo projection of pattern 1 (P1) in T5SS

Figure K.38 The Web-Logo projection of pattern 2 (P2) in T5SS

Figure K.39 The Web-Logo projection of pattern 3 (P3) in T5SS

Figure K.40 The Web-Logo projection of pattern 4 (P4) in T5SS

Figure K.41 The Web-Logo projection of pattern 5 (P5) in T5SS

Figure K.42 The Web-Logo projection of pattern 6 (P6) in T5SS

Figure K.43 The Web-Logo projection of pattern 7 (P7) in T5SS

Figure K.44 The Web-Logo projection of pattern 8 (P8) in T5SS

Appendix K.6 Web-Logo Projection of Type 6 Secretion Related Patterns

Figure K.45 The Web-Logo projection of pattern 1 (P1) in T6SS

Figure K.46 The Web-Logo projection of pattern 2 (P2) in T6SS

Figure K.47 The Web-Logo projection of pattern 3 (P3) in T6SS

Figure K.48 The Web-Logo projection of pattern 4 (P4) in T6SS

Figure K.49 The Web-Logo projection of pattern 5 (P5) in T6SS

APPENDIX L

REGEX FORMULA OF THE PATTERNS

Table L.1 RegEx formula of TISS

TISS Patterns	**Abrv.**
[IVALF]NXXGGXXG	P1
G[FSY][DF][IVA][DE][LIV][AME][KN][AE][IV]	P2
[LIAM][LATGSI][VALIT][GAFE][LVAID][STAISL][AGFL][TDPSGA][AYGL][ADLMPQF]	P3
[KRS][VIL][LKMVAIR][SAGKMQ][SALRPQ][KGNR][PA][DEV][AVFL][VIL]	P4
[KRPT][IVFL]G[VLAM][LVINYK][ALMVYPES][PDESTN][LMTV][STLV][GP]	P5
[NSQLTV][QNSREK][DETA][VCA][VADFIL][AGMTV][VILMA][VIFLM][G[PGSAH]	P6
[LVGTSIFMA]A{1,2}[LVGTSIFM]A{2,3}[LVGTSIFMA]A{1,2}	P7

Table L.2 RegEx formula of T2SS

T2SS Patterns	**Abrv.**
[GA][FMYL][TS][LM][LIV]E[VLMIAT][LMVIA][VLIAM][VAST]	P1
[LTAFVR][GAY][LYFAITV]G[PLRIKV][LIF][EQDN][PDAESV][LFMV]	P2
[VLACTQI][DE][AGVY][RTSFAH]L[PAKIR][DT]G[GS]R	P3
[IVFLM][GSTAV]G[GPASE][TMP][GAD][SAT]GK[TS]	P4
[DTEVS][LMAV][LFM][QTAVR][AV][ML][NSAQ][TS]GH	P5
[RLKVAHIQ][VILM][RK][YFLWIQ]R[IVCR]DG[VILEMQ][LM]	P6
G[PGAESTQR][TMVARS][GANS]G[KR][TS][TSVAIKLQ][TLSFMAV]	P7
[TS]G[HGR][LVIMP][MFVAIL] STA][TS][LIVMF]	P8
Q[RKV]L[VAIL][RQS][KRTAVP][LIVAR][CD]	P9

Table L.3 RegEx formula of T3SS

T3SS Patterns	**Abrv.**
.[LN]XX[LAGITV][LAY][LDQ][FLAQ][LS][GF][LAIVY]	P1
[VALG][IVAS]XL[LM][ILGY][AE][IDGL][YLVG][LI]	P2
[FIALR][PGC][SAG][LV][LGVM][LFK][IFSLM][TAR][TL]	P3
[ALV][AVDIL][GSADER][ALWY][LA][AFS]X[RAGQ][LAVH][AVLPT]	P4
DGAMKF[VI][KN]GD[TAS]	P5
[VL][VI][IFV][ILFV][AVI]T	P6
P[GLFMY][LAIV][GSA]X[ILMV][PD][LIVYAP][LIA][GP]	P7
[FILPR]X[LTAMSV][FLAW][AFLS][LV][LAV][LPIY][LPIY][LIAFGV]	P8

Table L.4 RegEx formula of T4SS

T4SS Patterns	**Abrv.**
[YFLWN][ERTSN]X[DN]X[ALV][GDHN][RNKQP][LVRMIPTK][TVISLANQ]	P1
[VI]X[QSTG]X[LMI][STM]L[RHN]L[KTV][DKN]V[PT]W[DKEQ]QAL	P2
L[RH]LX[DN]VPW[DKE]QAL[DEQA][ILTV][VI]	P3
[SKN][GD][AT][TAQ][ST][VIT][EST]F[KR][EK]A[AVM]L[GSA][LM][EKT]	P4
[FL][KREI][EK][AV][VAL]L[GSAE][LMT][ETK]VTP	P5
[GPVTFL][PGTS][TGP][GTALS][AGST][GKAS][KGSALT][STKVAL][TVASGT][TLASV][LTANG][LAYNGS][ALSYM][AMILT][LAI]	P6
[IV][PD][PAEDN][DEGRQ]ER[ILV][VIL][TC][IV]E[DE]	P7
GX[TS][LIM][ILVMS]E[LVMIT][MLIV]	P8
[GLS]X[TAV]X[YFW]X[YN][DN]XX[GLN]XX[TLS]	P9
P[DNG][IVMLR][IV][ML][VIL]GE[ILMAV]R[DGT]X[EDP]T	P10
[VYAL]X[SELV][LS][VAIL][SFV][LAIV]X[SK][LF]	P11
[EF][LI][LIVMF]X[LIVYMF][LIAVMF]XX[LIVMF]	P12

Table L.5 RegEx formula of T5SS

T5SS Patterns	**Abrv.**
GX[GDN][TS][LAIV][VTIL][LVGIF][SNTDR][SNTDR]	P1
[AILVG][STN][AGI][TSQNI][GDIN][LTD][VLAYGIT][ALIS][ALTGS][GNAY]	P2
A[STN][GA][VATGN][NSGQ][SAG][TSVIL][ASV][ILGVF]G	P3
[VALST][ANILV][LAVM][GT][ATGVPL][NTSG][ALNST]X[AVIW][TSPM]	P4
[TSN]DAV[NT][VLGT]XQ[LM]	P5
L[GNADK][DGANST][DE]X[SAQ][APLKV][TS]D[KRMHL][LVIM]X[ILV]X[GD]	P6
G[SGAL][YATV][LTAFGV]X[NSQL][LAINT]X[AMVE][AMFV][NQS]	P7
[AMV]A[NR][TNS][LMF]FXX[RQDS][LMAW][HR]	P8

Table L.6 RegEx formula of T6SS

T6SS Patterns	**Abrv.**
D[PR]X[VID][ES][RA][LMIV][LIF][EQ][GAS][FVR][ASC][FLGY]	P1
[VIPT][LVAIY]WXXGXX[LIVM]X[PT][QH][HL][FL]Q	P2
[LMV]X[CA][TVSA]P[AVIL][IVA]N[LAI]F	P3
[FL][LFAY][DN][ILFVAM][FY]X[HNT]R	P4
WXL[IV][SNR][HLNQ][LV][SNT][LFIM]	P5

APPENDIX M

RAW PATTERN SEARCH PERFORMANCE BY USING PSMS FOR DATASETS

Table M.1 Raw pattern search performance of TISS by using PSMS

TISS (954 protein CDHIT 0.5)	**5aa-Length**	**10aa-Length**	**15aa-Length**
Group1	59822	23430	9828
Group1 QC 5% Filtered	20438	902	214
Group1 QC 10% Filtered	4362	40	-
Group2	58028	20792	8174
Group3	60775	21845	7891
Group4	60038	21512	8244
Group5	58076	21153	8428
Total	318389	219648	102976
Total QC 10% Filtered	14805	174	1

Table M.2 Raw pattern search performance of T2SS by using PSMS

T2SS (668 Protein CDHIT 0.5)	**5-aa-Length**	**10-aa-Length**	**15-aa-Length**
Group1	42082	20792	8174
Group1 QC 5% Filtered	17757	1145	590
Group1 QC 10% Filtered	3883	154	60
Group2	47046	15719	7486
Group3	52118	15912	6529
Group4	42926	11379	4273
Group5	49416	17088	7908
Total	258132	153941	66050
Total QC 10% Filtered	18490	1119	533

Table M.3 Raw pattern search performance of T3SS by using PSMS

T3SS (381 Protein CDHIT 0.5)	**5-aa-Length**	**10-aa-Length**	**15-aa-Length**
Group1	15847	3309	1443
Group1 QC 5% Filtered	10672	937	431
Group1 QC 10% Filtered	2154	9	-
Group2	16816	3107	1143
Group3	19985	4229	1451
Group4	13393	2881	1438
Group5	18460	4347	2034
Total	110067	47092	18620
Total QC 5% Filtered	23171	346	105

Table M.4 Raw pattern search performance of T4SS by using PSMS

T4SS (770 protein CDHIT 0.5)	**5-aa-Length**	**10-aa-Length**	**15-aa-Length**
Group1	64374	14647	5074
Group 1 QC 5% Filtered	22680	621	348
Group 1 QC 10% Filtered	2489	4	-
Group2	62227	13840	4411
Group3	60642	14322	5828
Group4	63142	14485	4896
Group5	63036	13876	4500
Total	338393	186229	63893

Table M.5 Raw pattern search performance of T5SS by using PSMS

T5SS (221 Protein CDHIT 0.5)	**5-aa-Length**	**10-aa-Length**	**15-aa-Length**
Group1 QC 5% Filtered	55929	27399	9149
Group1 QC 10% Filtered	45814	8095	3186
Group2	43545	17176	7710
Group3	44280	19237	8479
Group4	58434	28136	11504
Group5	52038	24032	9042
Total	272686	208880	106739
Total QC 10% Filtered	165619	8162	1373

Table M.6 Raw pattern search performance of T6SS by using PSMS

T6SS (247 Protein CDHIT 0.5)	**5-aa-Length**	**10-aa-Length**	**15-aa-Length**
Group1	17059	4215	2863
Group1 QC 10% Filtered	3238	231	204
Group2	16744	3082	1868
Group3	13681	3405	2047
Group4	14114	2852	1770
Group5	15987	2908	1493
Total	105990	44723	25133
Total QC 10% Filtered	5518	126	60

APPENDIX N

FIVE-FOLD ASSAY RESULTS

Table N.1 Five-fold assay of TISS by using PSMS

TISS	Train (%80)	Test (%20)	# DS in TISS	T2SS	T3SS	T4SS	T5SS	T6SS
P1	120	39	117	0	0	2	2	0
P2	17	2	19	0	0	0	0	0
P3	49	11	60	41	17	28	7	3
P4	28	7	35	1	0	0	0	1
P5	30	12	42	1	0	0	0	0
P6	14	3	16	8	3	7	3	1
P7	46	15	57	26	16	26	5	0

Table N.2 Five-fold assay of T2SS by using PSMS

T2SS	**Train (%80)**	**Test (%20)**	**# DS in TISS**	**TISS**	**T3SS**	**T4SS**	**T5SS**	**T6SS**
p1	51	17	68			89	2	0
P2	18	5	23	0	0	4	0	0
P3	27	4	30	0	0	3	0	0
P4	78	14	92	0	0	21	0	1
P5	27	4	31	0	0	4	0	0
P6	25	5	30	0	0	5	0	0
P7	10	2	12	1	0	0	0	0
p8	43	7	50	0	0	10	0	0
p9	32	4	36	0	0	6	0	0

Table N.3 Five-fold assay of T3SS by using PSMS

T3SS	**Train (%80)**	**Test (%20)**	**# DS in TISS**	**TISS**	**T2SS**	**T4SS**	**T5SS**	**T6SS**
p1	7	1	8	0	0	0	0	0
P2	5	2	7	0	1	0	0	1
P3	12	3	15	1	0	1	2	0
P4	25	4	29	26	19	9	2	3
P5	9	3	12	0	0	1	1	0
P6	6	1	7	0	12	0	0	0
P7	22	4	27	0	27	7	1	0
p8	5	1	5	1	6	3	0	0

Table N.4 Five-fold assay of T4SS by using PSMS

T4SS	**Train (%80)**	**Test (%20)**	**# DS in TISS**	**TISS**	**T2SS**	**T3SS**	**T5SS**	**T6SS**
p1	97	23	68	0	1	2	0	0
P2	8	0	7	0	1	1	0	0
P3	11	1	12	0	2	2	0	0
P4	7	1	8	0	0	2	0	0
P5	12	1	13	0	0	3	0	0
P6	8	1	9	2	25	0	1	0
P7	6	3	9	0	6	0	0	0
p8	57	11	68	0	62	1	2	0
p9	46	3	49	0	0	0	0	0
p10	12	2	14	0	29	0	0	0
p11	7	1	8	0	0	0	1	0
p12	23	4	27	4	31	2	3	0

Table N.5 Five-fold assay of T5SS by using PSMS

T5SS	**Train (%80)**	**Test (%20)**	**# DS in TISS**	**TISS**	**T2SS**	**T3SS**	**T4SS**	**T6SS**
p1	87	20	44	5	6	2	12	3
P2	11	7	14	1	0	0	1	0
P3	34	6	20	1	0	0	1	1
P4	44	6	25	8	6	2	7	2
P5	80	13	55	0	0	0	1	0
P6	30	12	42	0	0	0	0	0
P7	6	6	12	1	2	0	0	0
p8	8	7	15	0	0	0	0	0

Table N.6 Five-fold assay of T6SS by using PSMS

T6SS	**Train (%80)**	**Test (%20)**	**# DS in TISS**	**TISS**	**T2SS**	**T3SS**	**T4SS**	**T5SS**
p1	16	7	23	0	0	0	0	0
P2	22	2	24	0	0	0	0	0
P3	28	8	28	0	0	0	0	0
P4	22	4	26	0	0	0	1	0
p5	13	4	17	0	0	0	0	0

APPENDIX O

PERFORMANCE OF THE PATTERNS IN IMMUNOGENIC DATABASE

Secretion Types	Patterns	Immunogenic Count		Secretion Types	Patterns	Immunogenic Count
TISS	p1	13		T4SS	p1	27
	P2	3			P2	0
	P3	105			P3	0
	P4	1			P4	0
	P5	6			P5	0
	P6	25			P6	0
	P7	124			P7	0
	Total	277			p8	9
					p9	5
T2SS	p1	1			p10	0
	P2	3			p11	7
	P3	3			p12	77
	P4	5			Total	125
	P5	1		T5SS	p1	18
	P6	0			P2	4
	P7	4			P3	1
	p8	2			P4	26
	p9	2			P5	4
	Total	21			P6	1
T3SS	p1	3			P7	7
	P2	2			p8	0
	P3	1			Total	61
	P4	26		T6SS	p1	0
	P5	0			P2	0
	P6	6			P3	0
	P7	3			P4	0
	p8	26			p5	0
	Total	67			Total	0

CURRICULUM VITAE

PERSONAL INFORMATION

Name, Surname	: Orhan ÖZCAN
Nationality	: Turkish
Date and Place of Birth	: 1981, Emet/Kütahya
Marital Status	: Married
Phone	: +90 262 677 3374
e-mail	: orhn.ozcn@hotmail.com

POSITION: Senior Researcher, THE SCIENTIFIC AND TECHNOLOGICAL RESEARCH COUNCIL OF TURKEY (TUBITAK) MAM, Genetic Engineering and Biotechnology Institute (GMBE). 41470 Gebze / KOCAELİ

EDUCATION/TRAINING
2005 B.S., Biological Sciences, Middle East Technical University (METU), Ankara, TURKEY
2008 M.S., Master in Biotechnology, METU, Ankara, TURKEY

POSITIONS
2006-2012 Research Assistantship. METU, ANKARA, TURKEY
2012-current Senior Researcher, TUBITAK MAM, GMBE, TURKEY

AWARD&HONOR
2012- 2214 TUBITAK Research award in abroad
2015 İlk İŞIM GIRISIM Semi-Finalist (Over than 300 project)

Research Experience
Comparison of the effects of nutritional and cultural conditions on lepidoptera-,diptera- and coleopltera- specific crystal protein production (2005-2008). As a researcher for proteomics studies

Preparation and optimization of high resolution proteome and sub proteome samples from Bordetella pertussis (2008-2011). As a researcher for proteomics studies

Comparative proteomic analysis of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica (2008-2009). As a researcher for proteomics studies

Improvement of protein subcellular localization prediction software (2009-2012). As a researcher for bioinformatics studies.

Development of Enzymes and Microorganisms for Efficient Hydrolysis of Agricultural Wastes and Lignocellulosic Raw Materials (2012 –current). As a researcher for Metagenomics studies.

Development of Hemostatic blood stopping Gauze

PUBLICATION & LECTURES & CONFERENCES
Orhan Ozcan, Bulent Icgen, Gulay Ozcengiz, 2010. Pretreatment of poultry litter improves Bacillus thuringiensis-based biopesticides production, Bioresource Technology. 101 (7), 2401-2404

Attend 2nd FEMS Congress of European Microbiologists (4-8 July 2006 Madrid- Spain) with a poster

Attend 15th National Biotechnology Congress (28-31 October 2007 Antalya- Turkey) with paper proposal

Attend XII İnternational Congress of Bacteriology and Applied Microbiology (5-9 August 2008 İstanbul-Turkey) with a poster

Production of Chitosan and Chitin &Chitosan by-products from 3 insect family. (Submitted to TUBITAK for Turkish and EU Patent Office)