BIOINFORMATIC ANALYSIS OF BACTERIAL AND EUKARYOTIC AMINO-TERMINAL SIGNAL PEPTIDES

CHOO KHAR HENG

(*B. Comp. (Hons.), NUS*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2009

Acknowledgements

Countless people have contributed in varying degrees to enable this work. My heartfelt appreciation goes to:

- Professor Barry Halliwell and Professor Fu Xin-Yuan for providing me the opportunity to undertake graduate studies at the Department of Biochemistry, National University of Singapore (NUS)
- Professor Shoba Ranganathan, my main supervisor. An opportune talk with her years ago catapulted me into the exciting world of biology. Her continual encouragement and guidance have been immensely helpful
- Co-supervisor, Dr. Tan Tin Wee who has guided me in many aspects pertaining to my candidature and career growth
- Dr. Martti T. Tammi, for giving me the opportunity to participate in his research group and interact with the members to exchange ideas
- Drs. Theresa Tan May Chin, Chua Kim Lee and Low Boon Chuan for granting me the opportunity to continue my pursuit of this candidature
- Dr. Ng See Kiong, my current boss at the Department of Data Mining, I^2R for his support and encouragement for me to tackle new projects while pursuing my candidature
- Drs. Christopher Baker, Kanagasabai Rajaraman and Vellaisamy Kuralmani for the numerous discussion and brainstorming sessions that we had and the resulting projects
- My collaborators whom I have the pleasure of working with, including Drs. Lisa Ng and Zhang Louxin

- My fellow graduate friends previously from the Bioinformatics Centre (BIC), NUS: Drs. Tong Joo Chuan, Bernett Lee Teck Kwong, Kong Lesheng, Paul Tan Thiam Joo and Vivek Gopalan. Lim Yun Ping for being such a wonderful friend
- Mark de Silva and Lim Kuan Siong for their unmatched assistance offered in IT services and the many tricks and tips that they have selflessly shared with me while I was at the Department of Biochemistry, NUS
- Staff at the Dean's office, Yong Loo Lin School of Medicine and the Department of Biochemistry, NUS for their help and prompt assistance in administrative matters, in particular, Fatihah bte. Ithnin, Maslinda bte. Supahat, Lim Ting Ting, Nurliana bte. Abdul Rahim and Musfirah bte. Musa
- The Nobel Committee for Physiology or Medicine, Karolinska Institutet, Sweden, for granting the permission to use certain images in this thesis
- Nancy Walker, Copyrights and Permissions Manager from the W. H. Freeman and Company/Worth Publishers, for granting the permission to use two images from the book "*Molecular Cell Biology 5th Edition*" by Lodish *et al.* in this thesis
- My endearing family members including my mother, grandma and my lovely '*Duude'* for their love, patience, support and encouragement

Table of Contents

Summary

Amino-terminal signal peptides (SPs) mediate the targeting of precursor secretory and membrane proteins to the correct subcellular compartments. Despite the availability of massive sequencing data in the past two decades, disproportionately little is known about their mechanism, targeting, excision and post-excision events.

To capture these sequences for creating a specialized and standardized resource for SP, we have developed a semi-automatic pipeline to extract SP-specific information from public sequence databases. 27,708 of the 356,194 sequences extracted from Swiss-Prot which purportedly contain SPs, were discovered to lack experimental support upon inspection. Consequently, "*SP filtering rules*" were formulated to systematically eliminate spurious and experimentally unsupported entries. Of the resulting 2,352 verified SPs, we were able to cluster and classify them into five major groups, including eukaryotes, Gram-positive and Gram-negative bacteria, archaea and viruses.

In analyzing the cleansed datasets, certain types of amino acid residues were observed to occur more frequently at specific positions in the vicinity of the SP cleavage site, as was previously suspected. However, the canonical "(-3,-1) rule" of (von Heijne, 1986a) which is based on the classical SP processing pathway, was found to account for only 61.6-77.5% of the total dataset. Non-canonical SPs appear to be devoid of standard sequence patterns. Yet, in the absence of a clear universal sequence motif, the entire process of protein targeting and excision occurs with remarkable precision, suggesting multiple mechanisms for SP recognition, as has now been verified experimentally by other groups. Most studies have hitherto focused on the primary structure of SPs, ignoring the possibility of structural features that may lie within this short peptide segment.

Therefore, to derive structural patterns in SPs, we developed a working structural model of the SP complex with its endogenous receptor through homology modeling, protein threading and structure compositing. Separate domains from crystal structures of *E. coli* receptor complexes were amalgamated to form a theoretical 3D computational model.

The model revealed various grooves that can only accommodate certain structural types of amino acid residues. The positions that these residues can occur, coincide with those observed at the sequence level. These findings inspired the development of a novel machine learning based prediction method.

Support Vector Machines were used to model both the structural spatial constraints and the linear sequence information. This approach, incorporating both canonical and non-canonical SP cleavage sites, has successfully predicted 80-97% of verified bacterial datasets in the benchmark against existing methods. Significative feature vectors were analysed and found to correlate with sequence positions, thereby providing structural support for the early use of the classical SP predictive rules. Structural grooves appear to be able to accommodate a variety of peptide structural motifs, including those that do not exhibit sequential patterns.

The successful use of structural features in this approach provides an explanation of the seemingly contradictory findings of site-directed mutagenesis studies such as Thornton *et al.*, 2006 and others, whereby sequence-based mutations gave rise to unpredictable SP processing outcomes. Hence, if structural data becomes available for eukaryotic SP, this approach may be useful for formulating more accurate methods and may be extendable to the prediction of other signal sequences.

List of Tables

List of Figures

List of Abbreviations

aa	Amino acid residues
ANN	Artificial neural networks
ATP	Adenosine triphosphate
B. subtilis	*Bacillus subtilis*
CaM	Calcium-binding protein calmodulin
cDNA	Complementary deoxyribonucleic acid
cTP	Chloroplast transit peptide
C-terminal	Carboxyl-terminal
DNA	Deoxyribonucleic acid
DOPE	Discrete Optimized Protein Energy
DsbA	Disulfide-bond A oxidoreductase
E. coli	*Escherichia coli*
EMBL	European Molecular Biology Laboratory
ER	Endoplasmic reticulum
Euk	Eukaryote(s)
FGF	Fibroblast growth factor
FN	False negative
FP	False positive
GO	Gene Ontology
GPCR	G protein-coupled receptor
Gram-	Gram-negative
Gram+	Gram-positive
GRAVY	Grand average of hydropathy
GTP	Guanosine triphosphate

GTPase	Guanosine triphosphatase
HGP	Human Genome Project
HIV-1	Human immunodeficiency virus-1
HLA-E	Human leukocyte antigen E
HMM	Hidden Markov model
ICM	Internal Coordinate Mechanics
MCC	Matthews' Correlation Coefficient
MHC	Major histocompatibility complex
MP	Mature peptide
mRNA	Messenger ribonucleic acid
MTS / mTP	Mitochondrial targeting signal / peptide
NES	Nuclear export signal
NLS	Nuclear localisation signal
NPY	Neuropeptide Y
N-terminal	Amino-terminal
Perl	Practical Extraction and Report Language
PDB	Protein Data Bank
p*I*	Isoelectric point
Preprotein	Precursor protein
Prl	Preprolactin
PTS	Peroximal targeting signal
RBF	Radial basis function
RNA	Ribonucleic acid
SARS	Severe acute respiratory syndrome
Sn	Sensitivity

SNP	Single nucleotide polymorphism
SP	Signal peptide
SPase I	Type I signal peptidase
Spc	Specificity
SPD	Secreted protein database
SPdb	Signal Peptide database
SPDI	Secreted Protein Discovery Initiative
SPF	SP fragment
SPP	Signal peptide peptidase
SR	Signal recognition particle receptor
SRP	Signal recognition particle
SVM	Support vector machines
Tat	Twin-arginine translocation
TN	True negative
TP	True positive
TrEMBL	Translated EMBL
UDP	Uridine diphosphate

Chapter 1: Introduction

1.1 Overview

The Human Genome Project (HGP) was initiated in 1990 with the primary aim of understanding the human genetic makeup. The project which spanned 13 years, identified over 20,000 genes with an estimated cost of USD300 million to sequence a human genome (the cost is estimated based on the parallel quest by Celera Genomics Inc.(http://www.genome.gov/11006943;http://ww.ornl.gov/sci/techresources/Human_Genome/home.shtml). Vast improvements in sequencing and high-throughput technologies since then, have made it possible to sequence a human genome under USD60,000 in less than a month (Applied Biosystems, 2008). Start-ups such as 23andMe or deCODEme Genetics are already capitalizing on the breakthrough to offer 'personalized genomics' services. They perform marker genotyping for individuals to learn about their own genetic profile and disease risk (Kaye, 2008). In January 2008, the "1000 Genomes Project" was launched to map the genomes of more than 1,000 individuals in an attempt to produce a detailed catalog of the genetic variations (http://www.1000genomes.org). These developments guarantee that the pace at which the sequence data are churned out will only accelerate.

The unprecedented availability of such voluminous data has literally transformed the study of biological and biomedical research. Now, it is a routine for experimental studies to involve informatic tools and computational techniques to collect, store, organize, retrieve, search, and to integrate the massive volume of sequence, structure, literature and other biological data from disparate data sources into a cohesive and coherent view for interpretation and analysis (Mount, 2001).

As the annotation of the immense data accruing from genome-scale projects continues to be an on-going ‘grand challenge’ for Bioinformatics and Computational Biology, assigning function accurately and effectively to the protein products encoded by the genes encapsulated in the genome sequences remains a significant barrier to our understanding of the functional molecules in cells (Louie *et al.*, 2008; Reed *et al.*, 2006). The role and function of a single protein depends on the partner proteins that it interacts with, which are in turn influenced by subcellular localization. Molecules secreted by a cell or an organism, often referred to as secretory proteins, play pivotal biological roles in the health and well being of an organism.

Secretory proteins reportedly represent 30% of the proteome of an organism (Skach, 2007) with functionally diverse classes of molecules such as cytokines, chemokines, hormones, digestive enzymes, antibodies, extracellular proteinases, morphogens, toxins and antimicrobial peptides. Some of these proteins are involved in a host of diverse and vital biological processes, including cell adhesion, cell migration, cell-cell communication, differentiation, proliferation, morphogenesis, survival and defense, virulence factors in bacteria and immune responses (Bonin-Debs *et al.*, 2004). Excretory-secretory proteins circulating throughout the body of an organism (e.g. in the extracellular space) are localized to or released from the cell surface, making them readily accessible to drugs and/or the immune system. These characteristics make these molecules as extremely attractive targets for novel vaccines and therapeutics, which are currently the focus of major drug discovery research programs (Bonin-Debs *et al.*, 2004; Serruto *et al.*, 2004). Several efforts have been carried out to accelerate the discovery of these proteins including the large-scale Secreted Protein Discovery Initiative (SPDI) which sought to discover novel secretory and transmembrane proteins in human (Clark *et al.*, 2003); identification of secreted proteins in 225 bacterial proteomes (Bendtsen *et al.*, 2005a) and the Human Proteome Folding Phase II (http://www.worldcommunitygrid.org/projects_showcase/viewHpf_2About.do). Such initiatives will likely increase with the completion of the numerous genome projects. These projects generate large number of novel sequences that require further annotations such as the identification of cleavable signal peptides (SPs) located at the amino-terminus of the secreted proteins as well as a subset of membrane proteins.

These SPs play critical roles in the secretory pathway where not only are they involved in targeting; they actually carry out additional functions post-cleavage processing. Surprisingly, we are only beginning to realize their tremendously diverse responsibilities as more studies continue to illuminate their functions (Hegde and Bernstein, 2006). This development has been somewhat disappointing especially when they have been discovered for more than three decades ago (von Heijne, 1998). One reason for this lack of interest is attributed to our unwarranted presumption that these peptides could not possibly possess much sophisticated functions beyond their short/small physique. Also, identification of SPs is often considered a secondary or lesser task of an experimental study. This is exacerbated by the relatively tedious effort required by experimental methods to identify the SPs, making them further unable to cope with the large influx of new sequencing data. Thus, *in silico* paradigm has emerged as a viable approach to complement traditional wet-lab experiments.

It enables specific studies to be carried out at a fraction of cost and time through simulation, prediction and others. Moreover, large-scale studies involving thousands of sequences concurrently are feasible and can be conducted relatively easier. Importantly, it allows for formulation of questions and testable hypotheses that are fundamentally different from traditional experiments, that otherwise could not have been developed with experimental approaches alone (Brusic, 2007).

1.2 Aims of Thesis

The goal of this thesis is to contribute to the understanding of the factors that govern the substrate specificity of SPs by means of bioinformatic and molecular modeling techniques. To attain this goal, the following objectives are established to:

I. Develop a robust and scalable pipeline for the generation and update of a high quality repository of SPs which shall form the foundation for subsequent undertakings of this work

II. Analyze the SPs sequences based on the dataset from (I)

III. Study the structure complexes of SPs to identify specific grooves that possibly could contribute the substrate specificity

IV. Develop a method for the accurate identification of the SPs cleavage site based on the insights obtained from (II) and (III)

V. Conduct a benchmark study using standardized dataset from (I) on the existing SP prediction tools and evaluate our newly developed method (IV)

While there is no lack of domain databases for the various types of sequence or structure data (http://www3.oup.co.uk/nar/database/c/), our survey showed that there was no specialized resource that catered to SPs when this work was initiated. Thus, the initial aim is to develop a customized pipeline to retrieve sequence entries from Swiss-Prot and extract selected information into a SP-centric repository. Maximal automation, ease of maintenance and scalability are set as important design criteria to cope with the continual deposition of new sequences.

Previous studies (Menne, *et al.*, 2000; Nielsen *et al.*, 1997) have highlighted the presence of erroneous annotations in the Swiss-Prot protein sequence database

(Bairoch *et al.*, 2004), but there was limited indication of the exact nature of the errors. It was also unclear the extent of the errors that was present. Hence, it will be useful to categorically classify these errors for formulating detection rules and techniques that could standardize the removal of affected entries. While identifying the errors, we want to explore the possibility of integrating information from nucleotide database - EMBL (Kulikova *et al.*, 2007) not only to augment the current repository, but also as an auxiliary method for error detection (Bork, 2000). Ultimately, these steps are to ensure that we can commence this work with a rigorously cleansed repository.

Next, we want to re-analyze the SP sequences including their amino acid composition, physico-chemical properties, which were investigated in previous studies (von Heijne, 1985; von Heijne, 1986a; von Heijne, 1986b von Heijne and Abrahmsen, 1989; Nielsen *et al.*, 1997), using our cleansed and enlarged dataset. In addition, we want to explore other properties such as isoelectric point, net charge, and to extend this exploration to the mature peptide (MP), which has received limited attention. The exploration of the MPs could help us to understand its influence and role in the cleavage event, in light of the report on its influence (Kajava *et al.*, 2000). Additionally, earlier studies have reported distinctive features that were exhibited by eukaryote, Gram-positive (Gram+) and Gram-negative (Gram-) bacteria groups (Nielsen *et al.*, 1997). It would be worthwhile to examine the basis for such distinction.

In these three groups of organism, their SPs were found often to be punctuated with an Ala-X-Ala sequence motif. The observation of the occurrences of this motif led to the formation of the ‘(-3, -1) rule’ (von Heijne, 1986a) which states that small and aliphatic residues are preferred at the -3 and -1 positions preceding the SP cleavage site. Some SP prediction tools have even incorporated this canonical motif as part of their rules in predicting the cleavage site (Gomi *et al.*, 2004). Since the proposal of this rule, more sequences have become available. Hence, the aim is to examine the validity of this rule and also to investigate possibly other non-canonical patterns that can be observable in the new sequences.

Most studies have largely focused on the primary structure of SPs. However, it has been reported that single residue substitution to the SP sequence is sufficient to cause a drastic effect (e.g. total abolishment in function or re-direction of targeting and so on) (Pidasheva *et al.*, 2005; Ronald *et al.*, 2008). While at other times, multiple substitutions or even deletion of a portion of the SP do not trigger any observable effect (Rusch *et al.*, 1994; Rusch *et al.*, 2002; Olczak and Olczak, 2006). We hypothesized that there may be structural features that lie within this short peptides. We want to study the structure of SP and its endogenous type I signal peptidase (SPase I) — the receptor enzyme that is responsible for the cleavage of SP from the mature peptide — for possible explanations to these observations.

However, there are currently four SPase I-substrate complexes that have been deposited into the Protein Data Bank (PDB) but they are of different substrates. If we extract selected domains from each of these structures as templates, the domains can be combined through computational techniques to develop a working model of the SP-SPase I complex. The knowledge gained from studying the SP-SPase I complex could cast a light on the propensity of certain residues to occur at specific positions as observed at the sequence level.

The combined insights from the analyses of SPs can be applied to develop new SP prediction method. There are two aspects involved in SP prediction: (i) detection of the presence of SP or in other words, to distinguish between secretory and non-secretory sequences; (ii) identification of the correct cleavage site. The aim is to develop a method that is able to tackle these two aspects by exploiting both the sequence and structural features. This could allow us to tackle non-canonical motifs as well. Following the development of our method, the next task is to benchmark the new method against other existing prediction methods using our standardized datasets. This will provide a fair comparison between the different prediction methods. The benchmark could help to establish if all the tools are able to perform equally well in both or just single aspect of SP prediction.

1.3 Thesis Organization

The rest of the thesis is organized as follows. *Chapter 2* provides a treatment on the background of SPs relating to their recognition and translocation machinery, interaction with the various partners in the early phase of the secretion pathway. To avoid any confusion, the usage of the terminology is standardized throughout this thesis. The unique characteristics and features of SPs are reviewed together with the cleavage processing mechanism. The post-targeting fate of the SPs is also described, followed by the presentation of the roles and functions of SPs. The chapter is concluded with a showcase of the applications of SPs in different domains.

Chapter 3 addresses the need for a high quality and centralized repository of SPs as an important prerequisite for sound analysis studies. The chapter details the methodology to develop a scalable bioinformatic pipeline capable of coping with new updates. The errors discovered in the collected public domain data are highlighted and solutions are proposed to tackle such issues. A short account of the developed system explains the system functions and features that are available for use.

Chapter 4 discusses the results from the large-scale computational analysis performed on SP-containing datasets. Various bioinformatic tools and techniques were applied to examine the different aspects of SPs including their primary sequence structure, sequence length and composition, physico-chemical properties and possible distinctive features around the cleavage-processing site. The MPs were also scrutinized in the study.

Chapter 5 describes the effort in generating the SP-SPase I-complex using 3D model constructed from the existing 3D structure data as a working model to understand the functional residues and the subsites involved in the substrate binding and specificity.

Chapter 6 presents the development of two SP prediction methods where the first is a matrix-based approach and the second describes a novel approach that differs from existing approaches by exploiting sequence and structural information. A brief review of the current state of prediction methods/tools is included, followed by a benchmark study of the existing SP prediction tools and the two newly developed methods.

The final chapter states the conclusion drawn from this work and summarizes the key contributions of this thesis to the advancement of understanding of SPs. Potential directions for future researches are suggested. The list of publications and presentations generated throughout the course of this work is included.

Chapter 2: Background on Signal Peptides (SPs)

Günter Blobel was awarded the 1999 Nobel Prize in Physiology or Medicine for his seminal work that "*proteins have intrinsic signals that govern their transport and localization in the cell*" (Blobel, 2000). This work was, in fact, initiated almost three decades ago. It was in 1971 when Blobel and Sabatini formulated the first version of "*signal hypothesis*" where they postulated the existence of a shared N-terminus sequence element among nascent polypeptide chain of secretory proteins (Blobel and Sabatini, 1971). The first experimental evidence in support of this N-terminus extension surfaced a year later when messenger RNA (mRNA) for the light chain of immunoglobulin G (IgG) was translated in a membrane-free translation system (Milstein *et al.*, 1972). Following this, an elegant *in vitro* coupled translation-translocation apparatus was developed to ascertain the function of this transient extension (Blobel and Dobberstein, 1975a; Blobel and Dobberstein, 1975b). The SP overall architecture was eventually elucidated with the availability of complementary DNA (cDNA) sequencing technology (von Heijne, 1983).

These landmark experiments formed the cornerstone for the discovery of other localization signals and paved the way for the design of various experiments in other biological systems. Genetic and biochemical studies followed to validate the "*signal hypothesis*" and confirmed the existence of such signal extensions in other preproteins including membrane proteins. A surge of interest in this emerging field ensued and these cumulative efforts have helped to advance our understanding of the individual components and pathways as well as the molecular mechanisms in cell, thus making a huge impact on modern cell biology.

Cells transport proteins to various intra- or extra-cellular locations such as endoplasmic reticulum (ER), nucleus and mitochondrial matrix, for insertion into a membrane or secretion out of the cell. This is achieved through a fundamental and important mechanism known as "*protein targeting*" or "*protein sorting*" (Pugsley, 1989). A myriad of proteins synthesized in the cell have to be transported into or across a membrane during their life cycle. This mission critical process requires timely and accurate export of proteins to their destinations by relying on the delivery information encapsulated in the short sequence segments known as "*signal peptides*" or "*targeting signals*" and the superb coordination of the translocation apparatuses (Dalbey and von Heijne, 2002). There are different classes of targeting signals that are involved in this active process of protein targeting, with each signal exerting their function in different cellular location (Figure 1).

2.1 Nomenclature of Targeting Signals

An impressive assortment of targeting signals exists in nature (see http://www.uniprot.org/docs/subcell *for the list of controlled vocabulary of subcellular locations and membrane topologies and orientations*). These targeting signals rely on specialized delivery mechanisms to be targeted the various organelles or cellular locations. These "*address labels*" or "*zip codes*" ensure that the passenger protein addressed to a specific destination is accurately delivered. There are also retention signals that anchor or confine the proteins to certain locations.

In general, these targeting or retention signals are located either at the ends (amino- or carboxyl-terminal) or they are embedded within the protein (internal). Different organelles are equipped with receptors that recognize and bind to specific type of signal sequence. The properties of the amino acids found in the signal region are likely to be important determinant in the interaction with the translocation machinery and the eventual destination of the protein. This was demonstrated in a proteomics and multivariate sequence analysis study, in which many of the experimentally identified proteins of *Synechocystis* with different physico-chemical properties in their SP and MP were routed to different extracytosolic compartments (Rajalahti *et al.*, 2007). Nevertheless, not all proteins possess a signal region; such proteins are usually retained in the cytoplasm. There is also a class of proteins that has a signal region but these proteins do not necessarily undergo cleavage processing.

A brief treatment of each type of signal here (Table 1) gives an overview to the multitude of targeting signals that has been discovered. The different targeted (sub)cellular locations are depicted in Figure 1. Two books have provided excellent reviews of these signals (Dalbey and von Heijne, 2002; Pugsley, 1989).

Table 1: Major classes of targeting signals are listed here with their targeted location. Each signal possesses its own unique characteristics and it is usually located at the N- or C-terminus of the preproteins. Motif patterns are represented using the PROSITE convention (de Castro *et al.*, 2006).

Signal name	Location	Features and description
Secretory / secretion signal / *Sec* signal / N-terminal SP	Endoplasmic reticulum (ER)	Located at the N-terminus of precursor secretory proteins. Possess the characteristic tri-partite structure where a hydrophobic core is conspicuous flanked by a positively charged n-region and a neutral, polar c-region. The cleavage site is located at the c-region. Uses the *Sec* translocation pathway to transport proteins in unfolded state (von Heijne, 1990).

Lipoprotein signal sequence	Cell membrane	Located at the N-terminus of bacterial lipoproteins and act as a retention signal. Similar tri-partite structure to secretion's n- and h-region but end with a lipobox which has the motif sequence [LVI]-[ASTVI]-[GAS]-C where a glyceride-fatty acid lipid anchor is attached to the Cys residue and cleaved by type II SPase (Tjalsma *et al.*, 1999) prior to the Cys residue. A PROSITE profile matrix is recorded for this signal (PROSITE Accession No.:PS51257).
Twin-arginine translocation (Tat) signal sequence	Membrane and periplasm	Uses the Tat pathway to transport protein in folded state instead of the Sec pathway. Similar overall design albeit with much longer length when compared with Sec signal. Notable differences include a consensus motif of [ST]-R-R-X-F-L-K motif (Berks, 1996) at the n-region; h-region has lower average hydrophobicity; positively charged residue in c-region with a Sec-avoidance motif (Bogsch *et al.*, 1997). Found in plants, bacteria and archaea.
Nuclear localisation signal (NLS)	Nucleus	Located either at the N-terminus or C-terminus. Nuclear proteins synthesized on free ribosomes in the cytoplasm are imported into the nucleus through a double lipid bilayer. Typically characterized by one or more clusters of basic amino acids (Hunter, 2007).
Nuclear export signal (NES)	Nucleus	Contrast to NLS, this is a signal for rapid nuclear export (Hunter, 2007).
Peroximal targeting signal 1 (PTS1)	Peroxisome	A trimer encoded at the C-terminal with the motif [SAC]-[KRH]-[LA] (Sacksteder and Gould, 2000).
Peroximal targeting signal 2 (PTS2)	Peroxisome	An N-terminus nonamer peptide with a consensus sequence [RK]-[LVI]-X(5)-[HQ]-[LAF] where X can be any amino acid residue. Less common as compared to PTS1 (Sacksteder and Gould, 2000).
Mitochondrial targeting signal / peptide (MTS / mTP)	Mitochondria matrix	Located at the N-terminus. Sequence is interspersed with alternating pattern of hydrophobic and positive-charge amino acid residues (Pfanner *et al.*, 1988; Schatz, 1993).

Chloroplast transit peptide (cTP)	Stroma	Located at the N-terminus. The sequence is rich in hydroxylated residues (Ser and Thr) but low occurrence of acid residues. A tri-partite domain is observed. Cleavage site is non-conserved although certain weak positional residues have been reported [IV]-X-[AC]↓A where (Emanuelsson *et al.*, 1999; Gavel and von Heijne, 1990).
Signal anchor	Transmembrane	Located at the N-terminus and act as a retention signal by anchoring the protein to the cell membrane. Often confused with N-terminus SP due to the presence of the hydrophobic domains (Martoglio and Dobberstein, 1998).
ER retention signal	Lumen	Located at the C-terminal and act as a retention signal by retaining the proteins in the ER lumen (Pugsley, 1989).
Signal patches	Nucleus	Uncleaved after sorting the protein from cytosol into the nucleus. Unlike other signals that are typically linear, locating these signals is non-trivial due to the non-contiguous manner in which they occur at the primary sequence but conjugated at the 3D dimensional space when the protein folds. NLS often exists in this form (Pugsley, 1989).

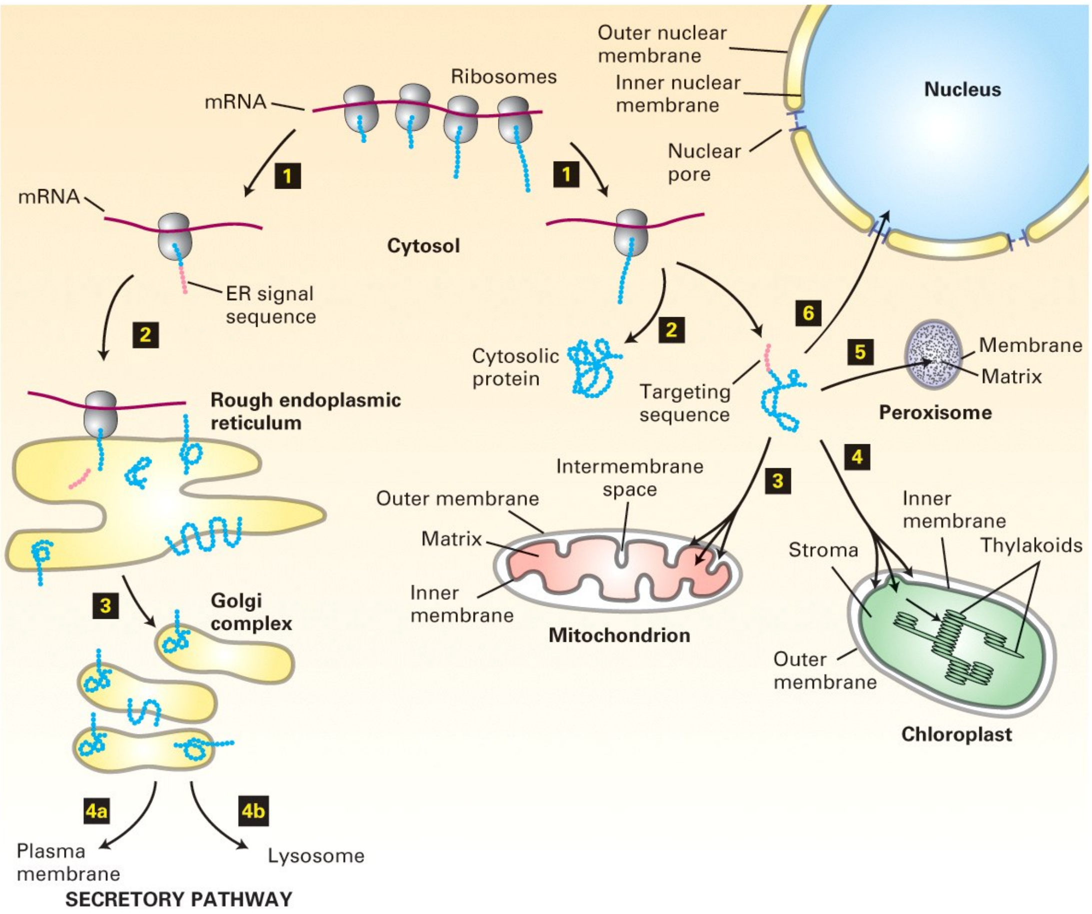


Figure 1: Schematic diagram of the various cell compartments in eukaryotic cell. The sequence in pink denotes the signal sequence whereas the blue sequence represents the mature protein sequence. This image is reproduced with permission courtesy of *W.H. Freeman and Company Worth Publishers* from the book *Lodish H., Berk A., Matsudaira P., Kaiser C. A., Krieger M., Scott M. P., Zipursky L. and Darnell J. 2004. Molecular Cell Biology, 5th Edition*.

2.2 Definition of SPs

One teething problem when a field such as this undergoes explosive growth is the uncontrolled use and introduction of vocabulary. Words or phrases are used interchangeably in a somewhat loose, ambiguous manner. Without a clear definition or agreement on a controlled set of vocabularies, confusion and miscommunication often follow. It is therefore crucial we provide a definition of the nomenclature used in this area of research to establish a common understanding.

Previous section introduces scores of targeting signals with each type of signal possessing its own unique characteristics. It is common to come across reference to these signals in the related literature as signal peptides, targeting signals, targeting sequences or signal sequences. Often, it is difficult to decipher the intended targeting signal without consulting the referred article. In particular, "*signal peptides*" is regularly used as a shorthand for the longer phrase "*N-terminus signal peptides*" — the most commonly studied type of signal — to refer to any of the targeting signal or simply as a generic term for all targeting signals. At times, it is used synonymously to describe "*leader sequences*" or "*leader peptides*" (Bowden *et al.*, 1992; Lam, *et al.*, 2003), even though they are of different nature and function. The state of misuse escalated to the point where there was a deliberate attempt to clarify on the usage of these terms (Molhoj and Degan, 2004).

In this thesis, we are particularly interested in the short *N-terminus signal peptides* of secretory proteins (comprise of mainly toxins, peptide hormones, digestive enzymes and antimicrobial peptides) as well as a subset of the single-pass type I membrane proteins where their N-terminal are exposed on the extracellular (or luminal) side of the membrane (Spiess, 1995). They mediate the targeting and translocation of the passenger protein domains across the ER membrane in eukaryotes or the inner and outer membranes in prokaryotes for insertion or secretion, upon which they are removed by the endoprotease SPase I (von Heijne, 1990; Spiess, 1995). Collectively, they will be referred to as "*signal peptide*" (SP) in this thesis to avoid repetitive mention of "*N-terminus SPs*". Our definition therefore omits signal sequences of lipoproteins, glycoproteins or other type I membrane proteins which are not cleaved by SPase I (Eichler *et al.*, 2003), including membrane proteins such as the mouse mammary tumor virus envelope protein and its alternative splice variant *Rem* which are also targeted to the ER but its signal sequence remains membrane-inserted (Dultz *et al.*, 2008). In case there is a need to refer to a particular type of signal, we shall specify the exact term according to the nomenclature (Table 1). "*Targeting signals*" or "*signal sequences*" shall refer to the different types of signals in general.

2.3 Characteristics of SPs

2.3.1 Overview

Secretory proteins are found in prokaryotic and eukaryotic cells where they are involved in a multitude of biological functions and processes. In human alone, approximately 30% of our proteins encoded by our genome are secreted or exported through the secretory pathway (Skach, 2007). Located at the N-terminus of these secretory proteins are short and transient polypeptides known as SPs which function as postal codes or address labels; they control the entry of virtually all proteins to the secretory pathway. Majority of these SPs are proteolytically cleaved during (co-) or after (post-) translation before eventually digested by peptidases (Figure 2). SPs are also found at the N-terminus of a subset of type I membrane proteins, particularly in eukaryotes though there were reports of their presence in other organisms as well, as we shall described in the later sections.

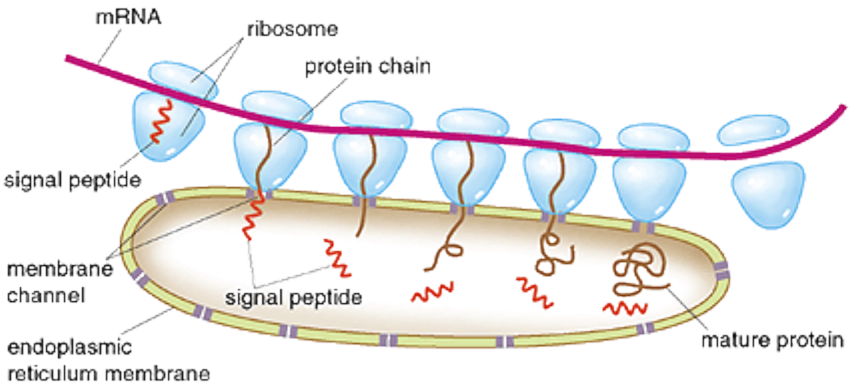


Figure 2: This simplified diagram shows a nascent polypeptide chain synthesized at the ribosome with a SP extension at the N-terminus. The SP directs the ribosome to the membrane channel of the rough endoplasmic reticulum and passes through the lumen and removed from the translating protein. The SP is absent from the mature protein. This image is reproduced with permission courtesy of the press release "*The Nobel Prize in Physiology or Medicine 1999*".

Comparative analysis of large number of known SPs across multiple species revealed limited homology. Nevertheless, these short peptides do possess common features and physical properties as well as some uniqueness. For instance, it was observed that there is higher incidence of Leu as compared to Ile in human SPs even though both possess similar hydrophobicity, though the bias was not detected in prokaryotes (Palazzo *et al.*, 2007). Interestingly, not all the features have to be present to qualify as a SP (Izard and Kendall, 1994). Functional SPs loosely conforming to these features have been reported and the variations purportedly augment the different modes in targeting and functions (Martoglio and Dobberstein, 1998). It is therefore not surprising when the SPase I has been suggested to recognize higher order structure rather than specific amino acids (pattern) at the cleavage site (Dalbey *et al.*, 1997). This could help explain the plasticity of eukaryotic and prokaryotic SPase I in recognizing each other's SP cleavage sites (Allet *et al.*, 1997; Osborne and Silhavy, 1993; Watts *et al.*, 1983).

The physical properties of the amino acids and features of SPs are important determinant in the interaction of the SPs with the various partners and in the localization of the protein within the translocation process. The SP-binding site at the SRP contains a large hydrophobic groove lined with Met residues, which supposedly confer the versatility to accommodate SPs of variable sequences and shapes due to the flexible side chains devoid of any branches (Keenan *et al.*, 1998). It was discovered in yeast cells that hydrophobicity ostensibly governed pathway selection; SPs of proteins that utilized SRP-independent pathway were found to be less hydrophobic than those that do not (Ng *et al.*, 1996). Such properties including charge, hydrophobicity and length, ensure that the SPs are properly interpreted to safeguard the accurate delivery of proteins their targeted destinations.

SPs generally have a short span of 13 to 36 amino acid residues (aa) though the average length varies with the organism groups (Molhoj and Degan, 2004). Prokaryotic SPs are generally longer than eukaryotic SPs (SP_{Euk}), in particular those belonging to Gram+ bacteria (SP_{Gram+}), which are usually 30aa long due to the longer h-region while SPGram-, are on average 23aa. SP_{Euk} are 22aa (Choo and Ranganathan, 2008). SPs with extended length have been reported, particularly those in bacteria or virus. Often, they are known to perform additional functions (Froeschke *et al.*, 2003). The shortest SP is found to be 11aa and the longest at 59aa in the SPdb (Albers, *et al.*, 1999; Choo and Ranganathan, 2005). A survey of literature reveals that the length of SPs can sometimes be extended without affecting its function albeit with lower efficiency. At other times, the extension may simply handicap the SPs (Pugsley, 1989).

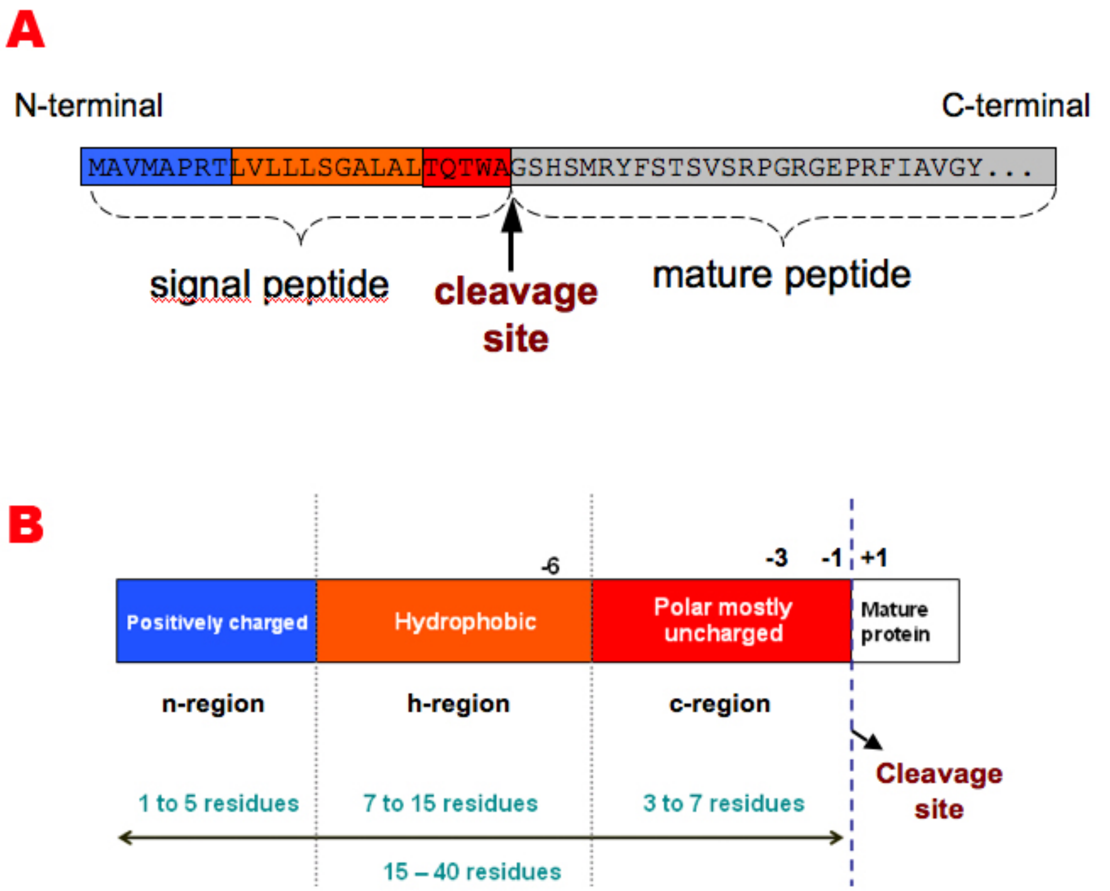


Figure 3: General architecture of a SP found in secretory proteins. (A) Cleavage site (blue dotted line) occurs at the interface of the signal and mature moieties. (B) An enlarged illustration of the SP that depicts the hallmark tri-partite structure. Cleavage occurs between the positions -1 (P1) and +1 (P1').

Figure 3 shows the general structural architecture of a SP sequence. A SP typically can be divided into three regions: (i) *h-region* is the hydrophobic core; (ii) n-region is located at the N-terminus and (iii) *c-region* is where the cleavage of the SP from the mature protein takes place. This "positive-hydrophobic-polar" architecture is thought to facilitate efficient binding to the lipid bilayers (von Heijne, 1990).

To standardize the conventions for addressing the different positions in the sequence, any position prior to the cleavage site shall be indicated as P1 (position -1), P2 (position -2) and so on hereinafter. For those positions after the cleavage site, they shall be indicated as P1' (position +1), P2' (position +2) and so on.

2.3.2 H-region – the central hydrophobic core

The hallmark feature of SPs is often described as having a tri-partite structure endowed with a central hydrophobic core, termed the "*h-region*" (Gierasch, 1989). The length of this core varies with organisms and it is usually lined with stretches of between 7 and 15 hydrophobic residues. Nevertheless, there are reports of unusually long hydrophobic core (relative to their homologous counterparts). An example is the SPs of *Xmrk* from the *Xipophorus* fish genus, a receptor tyrosine kinase that closely relate to the human epidermal growth factor receptor (Schartl *et al.*, 1998).

An early study described a non-uniform hydrophobicity profile for this *h-region*, with hydrophobicity peaking at the midpoint (von Heijne, 1982). Subsequent examination of *E. coli* preproteins suggested that the speed at which preproteins are processed correlates with the SP hydrophobicity. Lower limit of hydrophobicity saw preproteins being processed at a relatively slower pace, but it permitted membrane association and translocation whereas rapid processing of preproteins was observed in intermediate range of hydrophobicity. Beyond this level, insensitivity to transport inhibitors and substantial competition with the transport of other proteins happened. Thus, it was suggested that the increased hydrophobicity disrupted regulation and maintenance of the different secreted proteins. This theory possibly explains the 'non-optimal' hydrophobicity prevalent in SPs when they could have evolved to attain maximum hydrophobicity (Rusch *et al.*, 1994).

Another feature of this apolar region is its propensity to adopt α-helical conformation, particularly in a lipid or hydrophobic environment. Hence, this includes the case when it is bound to the signal recognition particle (SRP) (Plath *et al.*, 1998). Helix-breaking or turn-inducing residue such as Gly, Pro or Ser is commonly spotted at the downstream region (frequently at the P6 to P4) and they are often considered as the residues that demarcate the *h-* and *c-region* (von Heijne, 1990). These residues supposedly ease the insertion of SP through the membrane or translocation channel through the formation of hairpin-like structure (Driessen and van der Does, 2002), where the β-turn was suggested to facilitate catalytic processing of the SPase I cleavage site (Karamyshev *et al.*, 1998). Yamamoto *et al.* earlier investigated the significance of Pro residues at various positions (P10, P9, P7, P6, P5, P4 and P2) and found that secretion was impaired or lost when Pro was placed at different positions within the core (Yamamoto *et al.*, 1989). There were also studies that claimed the β-turn may not be a requirement; mutation or substitution of these residues that led to less efficient processing was attributed to reduction in overall hydrophobicity as opposed to conformational changes (Laforet and Kendall, 1991; Jain *et al.*, 1994).

The hydrophobic core is functionally crucial and it plays a critical role in allowing the SP to span across the bilayer membrane in eukaryotic or prokaryotic cells. It positions the SP strategically near to the lipid head group to facilitate cleavage, thus providing a plausible explanation to the failed cleavage when the hydrophobic core is extended beyond certain threshold (von Heijne, 1998). Also, hydrophobicity specifically the gradient within the core, as opposed to its overall hydrophobicity, is said to affect orientation (Goder and Spiess, 2003). Hydrophobicity supposedly influences the selection of the targeting route as well (Ng *et al.*, 1996), in addition to conformation of SPs (Zhen and Gierasch, 1996). Further, a point mutation study showed that this domain could conceivably influence the timing and efficiency of N-linked glycosylation and SP cleavage. The authors explored parameters including hydropathy, α-helical tendency or the Leu/Ile/Val and deemed that they are not the sole determinants. They suggested that other parameters may partake in regulating glycosylation efficiency, without ruling out the possibility that the information may be encoded in other manner as well (Rutkowski *et al.*, 2003). It was proposed that a threshold SRP-binding affinity might be necessary to enable translocation in yeast cells, and this is supposedly influenced by the hydrophobicity of the *h-region* (Bird *et al.*, 1987). Thus, mutations or deletion of even a single amino acid from this region has been shown to impair or abolish translocation activity, ostensibly disrupting the fine balance of hydrophobicity (Rusch *et al.*, 1994).

In essence, this region is sensitive to disruption, in particular with the introduction of charged or helix-breaking residue (Oliver, 1985). It has been reported that attaching a SP with sufficiently long stretches of hydrophobic residues can coerce a normally non-secreted protein to translocate to the ER lumen or inner membrane (Lodish *et al.*, 2004). This hydrophobic domain thus forms an important binding site that is critical for the translocation and targeting interaction and activity.

2.3.3 N-region – the positive-charged domain

Preceding or upstream of the hydrophobic core *h-region* is the "*n-region*", a net positive charge domain containing one or more Lys or Arg residues (von Heijne, 1990). This domain reportedly binds to the negatively charged phosphate group on the SRP 4.5S RNA (Batey *et al.*, 2000) and interacts with the ATPase SecA and negative-charge phospholipids in bacterial cells (Van Voorst and De Kruijff, 2000).

This domain typically contributes to the great variations in the overall length of SP (Martoglio and Dobberstein, 1998). The positively charged residues are evident in the bacterial SP, particularly in Gram-positive bacteria, but appear only sporadically in eukaryotic SPs. This apparent bias is possibly due to the formylated, uncharged N-terminal Met residue found in prokaryotic proteins as opposed to the unformylated, positively charged counterpart in eukaryotic proteins, thus compelling the former for the uptake of Lys or Arg as compensation (von Heijne, 1984b).

There have been indications that positive charge might influence (1) the efficiency of translocation where lesser net positive charge leads to slower rate in translocation (Izard and Kendall, 1994); (2) the orientation of the SP in the lipid bilayer (Spiess, 1995; Van Voorst and De Kruijff, 2000). Although there seem to be no explicit requirement on the positive charge in this domain, few studies have reported on the decrease in secretion efficiency may be due to influence of the positive charge in this domain (Gennity *et al.*, 1990; Guo *et al.*, 2008; von Heijne, 1990). It was also revealed that Levansucrase in *Bacillus* absolutely require positive charge in their SPs to direct secretion even though the net charge was negative, hence leading to the proposal that the presence of charge residues overrule the net charge as a requisite for a functional SP (Lammertyn and Anne, 1997).

In addition, the initial codons in the upstream of this region have been suggested to influence translational efficiency, particularly from the second codon to the fifth codon. Ahn *et al.* discovered that approximately 40% of *E. coli* SPs in their studies exhibit strong bias for the AAA triplet in their second codon. Similar high incidences of the triplet have been reported elsewhere. In their experiment, when the original codon was substituted with the triplet AAA, significant increase in expression level was observed whereas switching it to other triplets result in near complete abolishment (Ahn *et al.*, 2007).

2.3.4 C-region – proteolytic cleavage site

Located downstream of the hydrophobic core is the "*c-region*" which measures between 3 to 7aa in length and it is decorated with neutral, polar residues. In contrast to the *h-region*, this region adopts an extended β-conformation to facilitate easy recognition by SPase I (Karamyshev *et al.*, 1998).

This domain contains the proteolytic cleavage site recognized by the membrane-bound SPase I (Paetzel *et al.*, 2000). Small and neutral residues inclusive of Gly, Ser and Cys but predominantly Ala residues are preferred at P3 and P1'; these residues are thought to be critical clues for the recognition by SPase I, which led von Heijne to postulate the "(-3,-1) rule" (von Heijne, 1986a). The rule accepts that the residue at P1 must be small residues (Ala, Ser, Gly, Cys, Thr or Gln) but prohibits aromatic (Phe, His, Tyr, Trp), charged (Asp, Glu, Lys, Arg) or large polar (Asn, Gln) at P3. Further, Pro must be absent from P3 to P1'. Several studies have demonstrated that introducing or replacing the original residues of P3 to P1 may result in alternative cleavage sites (Fikes, *et al.*, 1990). It should also be noted that the region immediately after the *c-region* preferably should not contain charged residues such as Lys, Arg which might affect the secretion process (von Heijne, 1994).

Experiment data from a study into the limits of length variations of this *c-region*, with the introduction of minimal types of amino acids, indicated that the optimal length would be in the range of three to nine residues to promote efficient cleavage. Exceeding this range led to impaired processing or complete abolishment. The authors noted that exaggerated variation indeed occurred in this region, though these SPs are also unusual in other regards such as incredibly long *n-region*.

2.3.5 Mature peptide (MP) region

The peptide immediately after the cleavage site constitutes the MP where the passenger protein is subjected to further modifications such as formation of disulfide-bond, or addition of N-linked sugars and the likes before folding to a proper conformation to exert its function or targeted further elsewhere (Wollenberg and Simon, 2004).

Reflecting somewhat similar constraints as the other regions mentioned earlier, positive charged residues are not welcome, particularly at the N-terminus of this region in bacterial proteins. Neutral or net negative charge is favored in this region (Gierasch, 1989). *In vivo* and *in vitro* studies have reported deleterious effects upon the region in the presence of positively charged residues such as Arg or Lys. Nonetheless, the same does not apply to eukaryotes. This is perhaps due to the electrochemical potential across the inner membrane in bacteria where statistical analysis of membrane proteins have suggested that in prokaryotes, the cytoplasmic domain has generally more positive charge than the exoplasmic domain, thus giving rise to the “positive-inside rule” (Spiess, 1995; von Heijne, 1990).

2.4 Protein Synthesis and Cleavage Processing

2.4.1 Translation, targeting and translocation

Using the eukaryotic cell as an illustrative example, this section describes the protein synthesis and translocation processes and introduces the numerous main casts together with the ancillaries that interact with SPs along the pathway. The general concepts are somewhat related to other organism groups, though we shall describe some of the differences as well. A good understanding of these superbly orchestrated biological processes and the different molecular machineries involved will lay the foundation to appreciate the certain unique characteristics of SPs found in secretory proteins.

The synthesis process begins with the messenger ribonucleic acid (mRNA) carrying the genetic information from the DNA to the free ribosomes to be translated in the cytosol (Figure 4). The polypeptide chain can be translocated in two ways (Kalies and Hartmann, 1998):

(i) *co-translationally* — for secretory proteins translocating across the ER membrane, particularly those with more than 100aa. This is the most common route for the majority of secretory proteins. SP is recognized twice, with the first being recognized by the SRP and subsequently at the membrane (Rapoport *et al.*, 1996)

(ii) *post-translationally* — for smaller secretory proteins, certain yeast proteins, bacterial plasma membrane, mitochondrial, nucleus, chloroplasts and peroxisomes (Plath *et al.*, 1998). In yeast, SRP is reportedly required for efficient translocation though it is not essential for cell growth. The reliance on SRP-facilitated targeting of proteins thus becomes non-obligatory (Zheng and Gierasch, 1996).

In the impeccably-timed co-translational translocation (Figure 4), a cytosolic and rod-shape ribonucleoprotein complex termed "*SRP*" (Walter and Blobel, 1981a; Walter and Blobel, 1981b; Walter and Blobel, 1981c), consisting of six protein subunits of different molecular masses (termed "SRP9", "SRP14", "SRP19", "SRP54", "SRP68" and "SRP72") and a 300-nucleotide RNA molecule (termed "7SL RNA") (Walter and Blobel, 1982) swiftly binds to nascent chain complex (reviewed by Pool, 2005). Specifically, the Met-rich and conformationally flexible M-domain (Clemons *et al.*, 1999; Keenan *et al.*, 1998) or the NG domain (Cleverley and Gierasch, 2002) of the SRP54 subunit binds to the SP at the N-terminus of the nascent protein (Bernstein, 1998) as soon as the length of the nascent polypeptide reaches a certain threshold. The threshold is reported to be approximately 70aa out of which about 30aa are buried in the ribosome (Wiedmann *et al.*, 1987; Wollenberg and Simon, 2004). The binding reportedly triggers conformational change that activates SRP RNA (Bradshaw *et al.*, 2009).

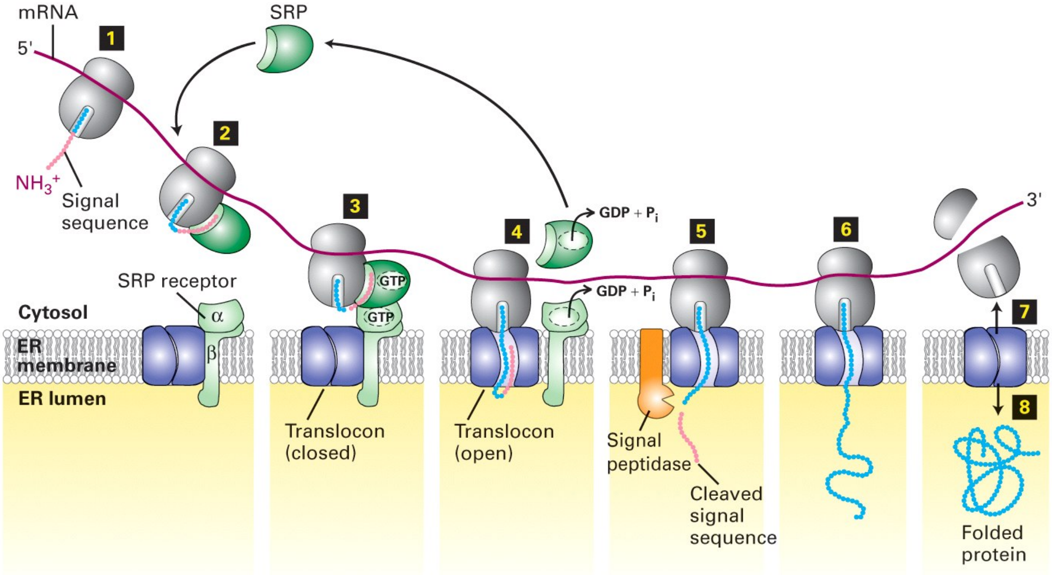


Figure 4: This diagram depicts the sequence where a protein is synthesized involving the translation of the nascent polypeptide chain to the cleavage processing of the SP (or known as signal sequence in the diagram) by the membrane-bound SPase I. This image is reproduced with permission courtesy of *W.H. Freeman and Company Worth Publishers* from the book *Lodish H., Berk A., Matsudaira P., Kaiser C. A., Krieger M., Scott M. P., Zipursky L. and Darnell J. 2004. Molecular Cell Biology, 5th Edition*.

The universally conserved SRP (*Ffh* or *fifty-four-homolog* is the bacterial homolog of SRP54) temporary arrests the translation of the polypeptide chain that is emerging from the ribosome. Although the elongation arrest is not compulsory, it is thought to promote efficient targeting by allowing sufficient time for proper placement of the ribosomes to the ER membrane (Walter and Johnson, 1994).

SRP then shuttles between the cytosol and the rough ER membrane to recruit the complex to a docking protein called "*SRP receptor*" (SR) that is situated at the rough ER membrane (Gilmore *et al.*, 1982a; Gilmore *et al.*, 1982b); this interaction cycle is mediated by the guanosine triphosphatase (GTPase) (Bradshaw *et al.*, 2009). The SR which exists as a heterodimer in eukaryotes (consists of an alpha-subunit peripheral membrane (SRα) and a beta-subunit transmembrane (SRβ) GTPases) or a monodimer in bacteria (FtsY being the homologue of SRα) (Gill and Salmond, 1990), then discharges SRP from the complex to permit the concomitant insertion of the SP+polypeptide chain through the dynamic protein conducting channel/pore known as *translocon* to resume (Walter and Lingappa, 1986). This disassociation is catalyzed by the SRP RNA where it accelerates GTP hydrolysis in the complex (Bradshaw *et al.*, 2007).

The translocon, termed "Sec61" in eukaryotes (Skach, 2007), is formed by three or four protein complexes of transmembrane proteins and estimated to be 40-60Å in diameter, a larger than expected size to maintain a permeability barrier and one of the largest holes observed in a membrane (Hamman *et al.*, 1997). It provides a sealed channel through the ER hydrophobic lipid bilayer and acts as a gatekeeper to control the passage into and out of the ER lumen (Crowley, *et al.* 1994; Romisch, 1999). The aqueous pore is gated on the lumenal side of the membrane and it is presumably closed by a lumenal protein such as BiP (Haigh and Johnson, 2002) that binds and blocks the pore. The pore is opened to the ER lumen only after the nascent chain reaches approximately 70aa in length (varies for different proteins) where the binding of the SP to the lumenal protein is reportedly the trigger for the opening of the aqueous pore. The length requirement is apparently critical as different studies on *preprolactin* (*prl*) have demonstrated that the extension to the length actually render a tightly sealed channel to attain translocation-competent state (Rutkowski *et al.*, 2001). This safeguards the mature region from being exposed to the cytosol (Crowley *et al.*, 1994; Hanein *et al.*, 1996). Other studies have similarly confirmed the interaction between SP and the translocon (Jungnickel and Rapoport, 1995; Mothes *et al.*, 1998). It is noteworthy that there are indicative differences in the manner with which individual SPs of different substrates initiate translocation and in the optimization steps involved for each protein, however, the extent of variance is currently unclear. Different SPs reportedly mediate early closure of the ribosome-translocon junction disparately (Rutkowski *et al.*, 2001).

Upon passing through the membrane, SP is excised from the growing polypeptide chain by the membrane-bound SPase I located on the lumenal or *trans* side of the membrane (Dalbey and von Heijne, 1992), with the C-terminus of the SP facing the lumenal side and the N-terminus orienting towards the cytosol (Goder and Spiess, 2003). A loop is temporary formed while the synthesis proceeds (see Figure 4 for illustration on the orientation and the reference *Goder and Spiess, 2003* for the various models proposed on protein topogenesis).

The elongation of the polypeptide continues until the protein is fully translated and the ribosome concomitantly dissociates from the ER. The polypeptide assisted by chaperones then folds into its proper 3D structure conformation to consummate the process before finally exerting their biological functions. Misfolded proteins are surrendered to degradation or ER-associated degradation (Crawshaw *et al.*, 2004). This example illustrates the case for soluble protein equipped with a "start-transfer" SP where the protein is synthesized and translocated in an N-to-C-terminal direction (Dalbey *et al.*, 1995) through the pore before settling into the ER lumen for further processing. In cases involving transmembrane protein, the protein similarly require a "stop-transfer" signal sequence that is uncleaved and effectively embeds the protein across the membrane (von Heijne, 1990). A recent review summarized the general principles of protein sorting in the secretory pathway (van Vliet *et al.*, 2003).

In bacteria, targeting of the nascent proteins can be accomplished through two other post-translational routes in addition to the co-translational SRP-dependent pathway just described, which essentially recognizes SPs with strong hydrophobicity. Two routes are utilized in light of failed recognition by the SRP where the first route entails the targeting of the preproteins directly to the translocase while the second involves a chaperone *SecB* which binds to long unfolded preprotein before binding to a peripheral subunit of the translocase, *SecA* (Driessen and van der Does, 2002). Several reviews have described in detail the translocation process and the related mechanisms in bacteria (Fekkes and Driessen, 1999; Holland, 2004; Harwood and Cranenburgh, 2008).

2.4.2 Cleavage processing by type I signal peptidase (SPase I)

The membrane-bound SPase I is responsible for the excision of SP from the growing polypeptide chain (for reviews, see Dalbey *et al.*, 1997; Ng *et al.*, 2007; Paetzel *et al.*, 2000; Paetzel *et al.*, 2002b; Tuteja *et al.*, 2005). This important cleavage event enables the liberated SPs to exert further biological functions. Current knowledge of these proteases are derived from examination of their sequences since only four crystal structures of SPases I have been resolved and they are all from *E. coli* (Paetzel *et al.*, 1998; Paetzel *et al.*, 2002a; Paetzel *et al.*, 2004; Luo *et al.*, 2009).

SPases I belong to a class of the serine protease family (Carlos *et al.*, 2000) and they are divided into 2 subfamilies (Tjalsma *et al.*, 1998; Ng *et al.*, 2007):

(i) prokaryotic(P)-type – bacteria, mitochondria and chloroplast;

(ii) eukaryotic endoplasmic reticulum (ER)-type – eukaryotic, archaeal and limited bacterial species

The P-type SPases reportedly exhibit substantial sequence similarity albeit differing in total length. Gram- SPases I (*E. coli*) are generally bigger in size than those of Gram+ (*B. subtilis*) though exceptions do occur where the latter is similar in size to the former (van Roosmalen *et al.*, 2004). In general, there are five regions or known as 'boxes' labeled from *A* to *E* (Dalbey *et al.*, 1997) which are conserved from bacteria to human with *Box A* being part of the anchoring domain and the rest involved in the catalytic mechanism in substrate cleavage (Ng *et al.*, 2007).

Although the substrate specificities of the Gram+ and Gram- SPases are known to be different, it remains unclear if it is related to their different characteristics in the SPs (*Chapter 4*). Interestingly, the catalytic Ser/Lys dyad retains its invariability across different bacterial SPases I (SPase I of *B. subtilis* chromosomal *SipW* reportedly uses Ser/His dyad (Paetzel *et al.*, 2000) and exhibit high degree of similarity to eukaryotic and archaeal SPases (van Roosmalen *et al.*, 2004). ER-type SPases largely utilize a catalytic Ser/His dyad in place of the Lys as observed in P-type SPases. The ER-type SPases are known to be much more complex (multimeric) than their bacterial counterparts. They are weakly homologous to the bacterial enzyme. In addition, unlike the active sites of bacteria that are easily accessible from the surface of the cytoplasmic membrane, the active sites of eukaryotes are buried within the ER lumen (Paetzel *et al.*, 2002b).

2.4.3 Post-translocation function and degradation of cleaved SPs

In spite of the improved understanding of the secretory machineries and mechanisms, our understanding of the fate of SPs upon its *coup de grâce* delivered by the SPase I remain limited. In eukaryotes, proteases involved in the further processing of SPs have yet to be characterized or discovered though the homologous counterparts are known in *E. coli* (Novak and Dev, 1988; Weihofen *et al.*, 2000). It is known that the remnant SPs excised from the mature protein are subjected to rapid degradation by the presenilin-type intramembrane-cleaving aspartic protease known as "*signal peptide peptidases*" (SPP) (Lemberg and Martoglio, 2002), giving rise to fragments which are released from the lipid bilayer to the ER lumen or to the cytosol (Lyko *et al.*, 1995; Martoglio *et al.*, 1997). It is notable that only a subset of SP substrates is also the substrates for SPP (Robakis *et al.*, 2008) even though SPP seem to be capable of catalyzing a wide variety of substrates including a viral protein in addition to the classical SPs (Martoglio and Golde, 2003b). The reason for this selective behavior is unknown. Other roles for SPP have been described including activation of signaling or regulatory molecules (Martoglio and Golde, 2003b), thus the roles of SPP in cell function could plausibly expand beyond degradation of SPs per se as we await further clarification.

Various studies have observed that the liberated SPs continue to serve important post-targeting biological roles (Jungnickel and Rapoport, 1995; Martoglio, 2003a). Early studies that examined the fate of the SPs upon cleavage certainly entertain that possibility. It was shown that freed SPs have to be cleared which might otherwise impede protein folding (Li *et al.*, 1996), and potentially having an impact on the subsequent functions in the secretion pathway (Koren *et al.*, 1983). Data from Martoglio *et al.* suggested that the SP fragments (SPFs), specifically the N-terminus moieties of the SPs from the hormone *prl* and the human immunodeficiency virus-1 p-gp160, possess regulatory function. When their SPFs were released into the cytosol upon proteolytic processing, they bound efficiently to the highly abundant calcium-binding protein calmodulin (CaM), which is known to regulate many protein targets in the Ca^{2+}-dependent signaling pathway, to antagonize Ca^{2+}-dependent phosphodiesterase *in vivo*, thus inhibiting CaM-dependent processes (Martoglio *et al.*, 1997; Martoglio and Dobberstein, 1998; Weihofen *et al.*, 2000).

Furthermore, subsequent evidence supported that hepatitis C virus was able to exploit the host's SPP processing and the series of cleavage events to aid in its protein processing towards maturation (McLauchlan *et al.*, 2002). Recent studies implicated the non-classical major histocompatibility complex (MHC) class I molecule human histocompatibility leukocyte antigen E (HLA-E) in the presentation of epitopes derived from the SPFs of MHC class I where the peptide-HLA-E complex interacts with the CD94/NKG2 receptors on the natural killer cells, thus wielding control over the functional activation and inhibition of the natural killer cells (Lemberg *et al.*, 2001). HLA-E surface expression is effectively influenced by the release of epitope-containing SPFs (Bland *et al.*, 2003; Braud *et al.*, 1997; Braud *et al.*, 1998; Lee *et al.*, 1998; Long, 1998). Similar studies had previously turned up with evidence that associated SPFs with antigen presentation (Henderson *et al.*, 1992; Hombach *et al.*, 1995; Wei and Cresswell, 1992).

These results implied that the SPs severed from the mature protein and the subsequent processing of them continue to wield material influence on the biological functions downstream of the secretory pathway or potentially other pathways including signal transduction pathways in the cell.

2.4.4 Non-classical signal sequences

Proteolytic processing of the secretory proteins is often necessary once the targeted destination is reached, to trigger the activation of subsequent events. A recent study on an essential protein involved in flagellum assembly called FliP, reinforced the need for the cleavage event. The motility function of *E. coli* was severely impaired when FliP was not cleaved (Pradel *et al.*, 2004). Nonetheless, not all secretory proteins possess signal sequences or are subjected to cleavage (Bowden *et al.*, 1992; Flower *et al.*, 1994), suggesting that other mechanisms or pathways for protein targeting exist. These proteins are termed "*non-classical*" secretory proteins. Some of these proteins are even known to have more than one function (Bendtsen *et al.*, 2005b).

Ovalbumin is a well-known example of a secretory protein that retains its signal sequence. The 100 residues N-terminus extension is found to be necessary for transport through the membrane to be effected (Tabe *et al.*, 1984). Serum paraoxonase/arylesterase 1 (Swiss-Prot ID: PON1_HUMAN) as well as the immunoevasin from the human cytomegalovirus US2 (Froeschke *et al.*, 2003) are some other examples. Another example is cyclophilin from the cattle parasite *Theileria parva*, which has a non-cleaved signal sequence that anchors the protein to the membrane upon targeted to the ER (Ebel *et al.*, 2004). However, in another cyclophilin found in *Drosophila rhodopsins* called ninaA, which has a membrane-spanning segment at the C-terminus, it was shown to possess a cleavable signal sequence (Stamnes *et al.*, 1991).

Ebel *et al.* had also earlier reported on p104 antigen also found in *Theileria parva* as being a non-cleaved protein (Ebel *et al.*, 1999). Cleavage of SPs usually occurs co-translationally, but there are instances where delay of the event occurs, for example in human cytomegalovirus US11 and HIV-1 glycoprotein 160 (Froeschke *et* *al.*, 2003; Rehm *et al.*, 2001). Another interesting find involved the G protein-coupled receptors (GPCRs) where one of the two groups requires the presence of cleavable SPs. The reason for the requirement is unclear though it was suggested that these SPs may aid in the translocation for those membrane proteins with impaired post-translational translocation (Kochl *et al.*, 2002). It was demonstrated that the presence of N-terminus cleavable SP is not essential in human hepatic membrane glycoprotein UDP-glucuronosyltransferase, which plays key role in drug metabolism since the protein was still targeted to the export apparatus (Ouzzine *et al.*, 1999). Such phenomenon of non-requisite of SPs was similarly observed in exported cell envelope proteins including alkaline phosphatase, β-lactamase, *MalE*, *LamB* and *MalS* which bore *prl* mutations, though the same did not extend to their cytoplasmic homologs (Prinz *et al.*, 1996).

There have been reports of proteins that do not possess signal sequence such as fibroblast growth factor 1 (FGF1), FGF2, *Engrailed* homeoproteins and interleukin1 (Joliot *et al.*, 1998; Bendtsen *et al.*, 2005b). Precursors of IL16 are yet another example without SP even though they are processed and secreted outside of cell (Baier *et al.*, 1997). These proteins do not utilize the classical secretory pathway and do not contain any characteristic motif; instead they are secreted through various non-classical pathways (Prudovsky *et al.*, 2003). Nonetheless, through methods such as amino acid composition, secondary structure and disordered regions, these secreted proteins could be identified (Bendtsen *et al.*, 2005b). Examples such as these will gradually surface as we hasten the pace of sequencing and discovery efforts.

2.5 Roles and Functions of SPs

SPs function like a postal address label on an envelope by mediating the transport of prokaryotic and eukaryotic secretory proteins to the ER for further processing. They are removed and degraded upon reaching the targeted locations, leaving them absent from the mature protein (Tuteja, 2005). Deletion of the SP such as from the ammonia channel protein *AmtB* in *E. coli* has been shown to cause dramatic reduction in *AmtB* activity due to the inefficient in translocation of the protein (Thornton *et al.*, 2006).

Long presumed as having the sole function of targeting the nascent chain to initiate interaction between the ribosome and the translocon, we have described in previous sections of the multiple roles that SPs carry to suggest otherwise. A growing body of evidences is affirming SPs deservingly possess far versatile functional repertoire (Hegde and Bernstein, 2006; Swanton and High, 2006).

It is known that SPs are involved in protein topogenesis (Spiess, 1995) and they reportedly stimulate the duration of translocation and regulate ribosome-translocon association in their post-targeting capacity (Rutkowski *et al.*, 2003). Further, SP serves as a ligand for the opening of translocation channel and additionally, manifests sequence-specific alteration on nascent polypeptide environment to attain favorable conformation (Rutkowski *et al.*, 2001).

In an early experiment involving designer-SPs, it was demonstrated that synthetic SPs which exhibit common structural features as original/authentic SPs, inhibited the processing of pre-prolactin, pre-forms of pancreatic digestive enzymes, and pre-placental lactogen. The SPs further prevented translocation of nascent chains when presented in high concentration (Austen *et al.*, 1984). Similar *in vitro* works further substantiated that free SPs indeed inhibited protein translocation (Chen *et al.*, 1987; Simon *et al.*, 1992) and modulated secretion (Koren *et al.*, 1983). A number of studies have found that SPs retarded folding of the mature part of the polypeptide (Li *et al.*, 1996; Park *et al.*, 1988; Weiss and Bassford, 1990) or affected polypeptide conformation (Oxender *et al.*, 1980; Roggenkamp *et al.*, 1985) as well as down-regulated gene expression (Serruto and Galeotti, 2004). Thus, SPs possibly influence the regulation of proteins to their destination (Kurys *et al.*, 2000; Li *et al.*, 1994).

In the experiment conducted by Briggs *et al.*, it was demonstrated that accumulated SPs might potentially impose deleterious effects on lipid bilayers (Briggs *et al.*, 1985). Additionally, several studies have separately established that yeast SPs demonstrate differential specificity in their pathway preferences as opposed to the mature proteins (Deshaies and Schekman, 1989; Feldheim and Schekman, 1994; Ng *et al.*, 1996). In a study on GPCRs, approximately 5–10% was shown to contain SPase I-cleavable SPs (Alken *et al.*, 2005). The SPs of this type of GPCRs were suggested to facilitate the expression of functional receptors and their presence was ostensibly dependent upon the features of their N-terminus (e.g. length, positive charges) (Kochl *et al.*, 2002). The reason for the additional SP is unknown other than being essential. In another study, the data indicated that rat SP of corticotropin-releasing factor receptor promoted an early step of receptor biogenesis (Alken *et al.*, 2005). In a study involving a T-cell receptor called cytotoxic T-lymphocyte antigen 4, a nonsynonymous polymorphism in the SP negatively regulates immune responses, and has been associated with risk for autoimmune disease. SP presumably determines the efficiency of post-translational modifications and the disease was attributed to inefficient processing of the autoimmunity (Anjos *et al.*, 2002).

Many other novel discoveries of SP functions have been documented of late. A foamy virus glycoprotein SP was shown to have a crucial role in viral assembly (Lindemann *et al.*, 2001) while SP was considered as an important factor in influencing viral infectivity with its involvement in lectin engagement (Marzi *et al.*, 2006). Interestingly, a study further substantiated that SP is perhaps capable of inducing protective immunity against a microbial pathogen *Coccidioides immitis* when administered as a gene vaccine or synthetic peptide. Previous reports have claimed that DNA vaccines were lower in efficacy with the omission of SP in their constructs (Jiang *et al.*, 2002). Hydrophobic fragments of SPs have been found bound to MHC complexes on the cell surface (O'Callaghan *et al.*, 1998) and more polar N-terminal fragments have been found bound to cytosolic calmodulin, implying possible signaling function (Martoglio *et al.*, 1997).

In *Section 2.4.3*, we described the fate of liberated SPs as a result of cleavage and subsequent processing events that involve the SPP (Martoglio and Dobberstein, 1997; Weihofen *et al.*, 2000). SPP was implicated in the generation of antigenic peptides from the SP of MHC class I molecules where the SPs of the corresponding proteins were suggested to exhibit regulatory function in immune surveillance of healthy cells (Lemberg *et al.*, 2001). Similar mechanism was likewise reported in hepatitis C virus where the virus hijacked the host's SPP processing and the series of cleavage events to marshal and prepare its proteins (McLauchlan *et al.*, 2002). Other post-targeting function such viral assembly have been reported (York *et al.*, 2004).

In addition, numerous studies have highlighted the adverse effects caused by mutation to SPs. Minor alteration or mutations to these SPs, even as slight as a single amino acid substitution or the lack of SP have been implicated in the onset of a number of diseases and complications (Chou, 2001b; Nielsen *et al.*, 1997). A missense mutation in the hydrophobic core of SP detected in half of the patients rendered a non-functional COL5A1 which encodes for a type V collagen culminated in the Classic Ehlers-Danlos syndrome, a heritable connective tissue disease characterized by skin hyperextensibility, atrophic scarring, joint hypermobility and generalized tissue fragility (Symoens *et al.*, 2008). Similarly, it was discovered that a single mutation of Cys to Arg in the hydrophobic core of its SP of human preproparathyroid hormone is enough to cause autosomal ominant familial isolated hypoparathyroidism where the mutation impairs secretion of the hormone (Datta *et al.*, 2007). Scores of other human inherited disorders have been associated with SPs arising from mutation, including familial central diabetes insipidus (Ito *et al.*, 1993), coagulation factor X deficiency (Racchi *et al.*, 1993), Schmid metaphyseal chondrodysplasia (Chan *et al.*, 2001), dentine dysplasia type II (Rajpar *et al.*, 2002), neurohypophyseal diabetes insipidus (Rittig *et al.*, 2002), thyroxine-binding globulin deficiency (Fingerhut *et al.*, 2004), familial hypocalciuric hypercalcemia (Pidasheva *et al.*, 2005), autosomal dominant hereditary pancreatitis (Kiraly *et al.*, 2007) and Weill-Marchesani syndrome (Kutz *et al.*, 2008). A recent study of type I diabetic patients revealed a novel mutation in the preproinsulin SP where it was linked to diabetes onset (Bonfanti *et al.*, 2009). A more surprising discovery reported findings on a body-weight regulation protein called Neuropeptide Y (NPY), which controls food intake and energy balance. An SNP in the SP of that secretory protein potentiated NPY-induced food intake (Ding, *et al.*, 2005). A list of SNPs-related disorders not described here can be found in (Jarjanazi *et al.*, 2008).

The accumulating findings suggest that the SPs and their subsequently liberation from the MP (with the ensuing processing of them) may have substantially far-reaching implications. SPs surely warrant further investigation of their properties and their neighboring residues to advance our understanding of SPs for their crucial roles in the secretory pathways of both prokaryotes and eukaryotes.

2.6 Surprising Complexity of SPs

SPs with their deceivingly short sequence and lifespan have led many to relegate them as simple and unsophisticated peptides. However, the notion of multi-faceted roles for SPs that outgrow their sole protein targeting function is fast retiring this false misconception with the growing body of evidence elucidating their true diversities and complexity (Hegde and Bernstein, 2006).

It has long been observed of the mutual recognition of SPs in bacteria and eukaryotes by certain conserved translocation components (Osborne and Silhavy, 1993). SPs supposedly can be swapped between different proteins without loss of their targeting functions (Izard and Kendall, 1994; Belin *et al.*, 2004). Further, it was established that attaching a SP to a protein through recombinant DNA technique was sufficient to direct a chimeric protein to translocate to the ER to be secreted even though the protein was originally devoid of such a sequence (Burghaus and Lingelbach, 2001). In fact, an earlier study involving the combinatorial swaps of yeast invertase SPs with seemingly random peptides estimated that 20% of the 'pseudo-signals' were functional, or at least partially functional (Kaiser *et al.*, 1987). Even so, there is no lack of studies that refuted these claims (Al-Qahtani *et al.*, 1998).

These findings raise a series of questions. The heterogeneity of SPs of different secretory proteins is well documented and they are known to share few similarities in their primary structure. Yet, in regard to the 'pseudo signals' study, the seemingly random peptides that were generated to be functional press the question: what is the permissible extent of variability for a SP before it becomes dysfunctional? Are SPs really all that similar? Are SPs admittedly as flexible and tolerant as expected? If interchangeability is really viable, why is there a need to devise the huge diversity when much simpler variations of SPs could have ostensibly accomplished the tasks? Wouldn't such diversity have succumbed to evolution pressure? Additionally, has Occam's Razor been overruled in favor of plurality in the case of the notorious sequence diversity long observed in SPs? How do the components of the machinery in the secretion pathway cope with the degenerate feature of SPs while simultaneously maintaining the high specificity and high fidelity requirements in the targeting, recognition and cleavage of SPs? Also, if SPs were indeed as multi-faceted as suggested, how can all the necessary information for carrying out their function be practically encoded within the short peptide length without escalating complexity further?

Growing body of evidence is challenging the dogma that SPs are functionally equivalent and mostly interchangeable (Bird *et al.*, 1987; Jungnickel and Rapoport, 1995; Kang *et al.*, 2006; Rapoport *et al.*, 1996). An early review aptly summarized previous findings in support of this (Zheng and Gierasch, 1996) where a quoted example described the failure of the SP of yeast carboxypeptidase Y to direct the export of its passenger protein in mammalian cells, incongruent to what has been reported for most precursors of yeast proteins. Translocation was only achievable when the SP of the yeast was modified or a mammalian SP was used (Bird *et al.*, 1987). Exchange of SPs has been shown to decimate virion infectivity (Pfeiffer *et al.*, 2006; Weltman *et al.*, 2007). Site-directed mutagenesis studies on SPs or SPs fused with heterologous proteins (including reporter) have only continued to corroborate that SPs are not as amenable as assumed, albeit at times, conflicting reports from different studies of the same SP subject added further confusion (Belin *et al.*, 2004; Blanco *et al.*, 1999; Frate *et al.*, 2000; Gennity *et al.*, 1990; Izard *et al.*, 1994; Kaiser *et al.*, 1987; Thornton *et al.*, 2006). It must be noted that often, a limited number of SPs were investigated in these studies and they may inadvertently over-generalize.

Additionally, site-directed mutagenesis experiments study mutations by replacing residues that are different in terms of properties from the originals, thus the substitution(s) may affect the overall protein configuration (Pugsley, 1989). This is congruent to the data which indicated that replacement of the residues might have altered the conformation/placement of the SP and the nascent chain critical for proper protein biogenesis (Rutkowski *et al.*, 2001).

An early study on *B. subtilis* described the different combinations of SP and MP that might have influenced protein secretion efficiency (Himeno *et al.*, 1986). Similar sentiment was echoed in the study undertaken by Kim *et al.* to investigate a set of SPs from different substrates through the translocation process. They observed a broad range of SPs with varying efficiencies in initiating translocation and proposed that the link between the SP and the MP is possibly interlocked (Kim *et al.*, 2002), which we termed herein as "*SP-MP coupling*" theory. A systematic screening of *B. subtilis* SPs further reinforced the claim of an optimal-fit between a given SP and its respective MP (Brockmeier *et al.*, 2006). This relationship could ostensibly aid in explaining the conservation of SPs across species for a given substrate in contrast to the SPs sequence divergence observed for different substrates (von Heijne, 1985; Williams *et al.*, 2000). Collectively, post-targeting functional differences between SPs, substrate-specific evolutionary conservation and the adverse effects on altering SPs have been suggested as plausible explanation for the sequence diversity observed in SPs (Kim *et al.*, 2002).

Substrate-selective modulation of protein translocation was also suggested as the rationale behind the sequence variability, thus reconciling the seemingly paradoxical existence of 'imperfect' SPs such as those from the Prolactin hormone which would be detrimental to the cell under certain cellular conditions when they could have evolved to efficient ones (Kang *et al.*, 2006). Similarly, an earlier study observed that many proteins did not use optimized SPs in their targeting process (Levine *et al.*, 2005). Counter-intuitive as this may seem, such design certainly connote a functional intent rather than a random variation. This essentially equipped the cell to selectively modulate the release of certain proteins depending on particular conditions such as during ER stress where cargo proteins are barred from entering the ER while molecular chaperone BiP is permitted (Kang *et al.*, 2006), we likened this to a 'knob or tuner' for the cell to regulate the release of proteins on a demand basis.

Exceptions to the widespread view of SPs sequence diversity do occur. For instance, the SPs of conotoxin found in a small disulfide-rich peptide from the predatory cone snails that targets components of neuromuscular system, manifest hyper-conservation of SPs (Olivera *et al.*, 1999; Wang and Chi, 2004). Another example is the translocation-efficient caseins, which accounts for 82% of proteins in bovine milk (Creuzenet *et al.*, 1997; Watson, 1984). But such extreme examples are few. To cater to the large variety of secretory proteins with different functional and physiological requirements, we opined that this strategy is seemingly out of necessity and hinted at a far more sophisticated and complex mechanism at work.

2.7 Relevance and Importance of SPs

The pioneering works of Blobel's and other discoverers have revolutionized many aspects of modern cell biological research and set the tone for a blazing pace of research in this area. Research into the biology of prokaryotic and eukaryotic cells, advancement in molecular genetic techniques, improvements in large-scale cultivation and production of heterologous proteins have opened up new avenues and compelled researchers to actively pursue development in the protein targeting domain. With the majority of the proteins synthesized in the cytoplasm, these proteins have to be exported to the correct targeted destination to carry out their functions under the directive of SPs. Commanding a good understanding of SPs and the related mechanisms will unleash and broaden commercial applications in the pharmaceutical, medicine, biology, food industry and other areas.

The huge demand from the growing worldwide population is straining the limited natural resources and urgently calls for the advancement in gene technology. Designing recombinant DNA sequences that are highly optimized and efficient in secretion of heterologous gene products is thus sought after as a lucrative and desirable solution. This can be achieved by fusing a SP to the desired protein, where the recombinant protein can be delivered to a desired location in heterologous production hosts such as *E. coli* or *B. subtilis* to be harvested or identified (Westers *et al.*, 2004; Harwood and Cranenburgh, 2008). One such example involved the use of the cleverly crafted "*phage display*" method (Rosander *et al.*, 2002) for the pathogenesis study in various bacteria such as *Staphylococcus aureus* (Rosander *et al.*, 2002) and *Streptococcus equi* (Karlström *et al.*, 2004). The fusion proteins attached with SPs are targeted to the cell membrane to facilitate easy isolation and characterization. Another technique called *Signal-exon trap* took advantage of the presence of SP in secretory or membrane proteins to devise a detection technique to identify such proteins on genomic scale. Such technique could be used for detecting chimeric proteins, growth factors, receptors, matrix-binding proteins and so on apart from natural proteins (Chen and Leder, 1999; Péterfy *et al.*, 2000). Protein drugs such as growth hormone, insulin, interferon to name a few, have been engineered for extracellular secretion to facilitate easy purification. Such efforts could be further elevated with programmable microprocessor implants for timed release of the substances (Langer, 1998; Santini *et al.*, 1999). In other attempts, non-secretory proteins were attached with targeting signals and secreted to the extracellular, common in the biotech protein engineering application (von Heijne *et al.*, 1994).

Bacterial cells encoded with the recombinant genes are often employed as protein factories, due to the ease of handling and growing. Nonetheless, producing human proteins often require more complex cell systems such as yeast cells in order to be functional. Bacteria and eukaryotes may share similar or equivalent counterparts in their translocation and secretion machineries and components, however, SPs from one organism generally do not function efficiently or greatly diminished when placed in another host. In certain situation, the activity may be entirely lost or changed. Various efforts have identified suitable SPs, host proteins and expression systems (Brockmeier *et al.*, 2006; Jacobs *et al.*, 1997; Lal *et al.*, 1999; Le Loir *et al.*, 2005; Nene and Bishop, 2001; Nouaille *et al.*, 2005; Olczak and Olczak, 2006; Schaaf *et al.*, 2005; Tan *et al.*, 2002; Yamamoto *et al.*, 1987). Another line of work involves tweaking secretory SPs to achieve higher efficiency in secretion (Barash 2002; Bardy *et al.*, 2005) (review specifically related to *Streptomyces* (Lammertyn and Anne, 1998) and *Lactococcus lactis* (Ravn *et al.*, 2003)). There were also experimentations involving various SPs in search of efficient ones to assist in the development of bacteria as vaccine carrier (Wu and Chung, 2006). Such developments are particularly critical and useful as secretory proteins are heavily employed in protein therapeutics, for example, one could reprogram cells for gene therapy purpose (Grabley and Thiericke, 1999).

In the attempt to raise crop yield, for instance, gene conferring herbicide-resistance was introduced into a tumor-inducing bacterium *Agrobacterium* *tumefaciens* where it is fused with a transit peptide and targeted to the chloroplast (Della-Cioppa *et al.*, 1987), thereby producing plants which are resistant to herbicide. Such technique can be expanded to include other herbicides and making the plants unharmed by the toxic substance. Similarly, albeit to a different targeted destination, Asayama *et al.* designed a new class of antioxidant comprising manganese porphyrin and a mitochondrial targeting signal (Asayama *et al.*, 2006).

For the same reason that attaching a SP can direct the passenger protein to the desired location, a defective or mutant SP can lead protein astray, for which they have been implicated in the onset of scores of genetic diseases including cystic fibrosis, familial hypercholesterolemia (result in high low-density lipoprotein). Another example involves a rare metabolic disorder called "*primary hyperoxaluria type I*" where the development of kidney stones at an early stage is caused by a mislocalization of a peroxisomal glyoxylate aminotransferase to the mitochondrion, culminating in overproduction oxalate due to the enzymatic deficiency to prevent accumulation (Purdue *et al.*, 1991) (refer to Section 2.5 for more examples). Thus, correcting such defects through gene therapy could present as another viable treatment regimen. Also, newer vaccine strategies increasingly seek to pinpoint constituents in microorganism that are recognizable by the cellular immune system and concentrate production of vaccine targeting such regions (Buus 1999; Corradin and Demotz, 1997). Efforts are also directed towards the interaction partners of SP such as SPases, which play essential roles in the viability of bacteria (Date 1983; Klug *et al.*, 1997), making SPases attractive targets for the design of novel antibiotics (Black and Bruton, 1998).

Chapter 3: Construction of a High-quality SP Repository

3.1 Introduction

Commencing a study with high quality dataset is crucial and demands equal care and rigor as other activities, especially in a bioinformatic/computational study. Any data bias, errors or incompleteness that is present or inadvertently introduced will likely render subsequent analysis, inference or conclusion unreliable. The consequence is particularly acute in the development of SP prediction tools, in which the datasets are used to construct the guiding models. Noise and errors in the datasets can be detrimental to the construction of the predictive models (Nielsen and Krogh, 1998). The accuracy of a predictive model is therefore highly dependent on the quality of the datasets, and may affect the constructed model in generalizing to new dataset.

As a result, researchers often devote considerable time sifting through primary databases such as Swiss-Prot (Bairoch *et al.*, 2004), EMBL (Kulikova *et al.*, 2006) and the likes to collate and construct specific subsets of these datasets. This repetitive process can and should be eliminated with the creation of a centralized repository. With a shared resource, benchmarking can be conducted in a standardized manner, unlike current situation where comparison between SP prediction tools is often difficult, or impossible due to the varied datasets used.

The website of the Nucleic acids research journal maintains a catalog of some of the databases that have been published (http://www3.oup.co.uk/nar/database/subcat/3/7/). Several resources exist that capture specific information on protein subcellular localization (http://www.bioinfo.tsinghua.edu.cn/~guotao; http://npd.hgu.mrc.ac.uk/; http://www3.oup.co.uk/nar/database/subcat/3/7/),nuclear proteins (http://npd.hgu.mrc.ac.uk) and secreted proteins (http://spd.cbi.pku.edu.cn/spd_index.php). These specialist databases do not provide SP-specific information except for secreted protein database (SPD) (Chen *et al.*, 2005), which assembles human, mouse and rat protein sequences from databases such as *Tr*anslated *EMBL* (TrEMBL) (Bairoch *et al.*, 2005), Ensembl (Birney *et al.*, 2004) and Refseq (Pruitt *et al.*, 2005). Datasets from the SPDI (Clark *et al.*, 2003), a large-scale effort to identify novel human secreted and transmembrane proteins; the Riken mouse secretome and seven other related datasets (http://spd.cbi.pku.edu.cn/help/spd_help.php) are found in SPD as well. SPD aims to be a comprehensive repository for secreted proteins, but it suffers from poor data quality due to the underlying data sources such as TrEMBL that it uses. TrEMBL, for instance, contains sequences generated from an automated protocol that have yet to be manually curated, and its SPs are predicted using computational methods. Furthermore, entries in SPD were not manually checked against publications. The lack of updates has made it obsolete. A closely related database is the Hera database that aggregates human ER proteins (include transmembrane proteins) from various protein sequence databases. In this database, the SPs of ER proteins are again predicted using computational tool (Scott *et al.*, 2004).

Besides specialist databases such as SPD that offers datasets for download, several websites offer SP datasets as well (Menne *et al.*, 2000; Nielsen *et al.*, 1997). One of the earliest efforts was the compilation of 277 targeting signals that included SPs (Watson, 1984). In recent times, there is the dataset of 270 secreted recombinant human proteins with experimentally determined cleavage sites (Zhang and Henzel, 2004). However, these datasets are either limited in size or otherwise lacking in tools for querying the datasets. Some are outdated, as they do not keep updated with the publicly accessible primary sequence databases.

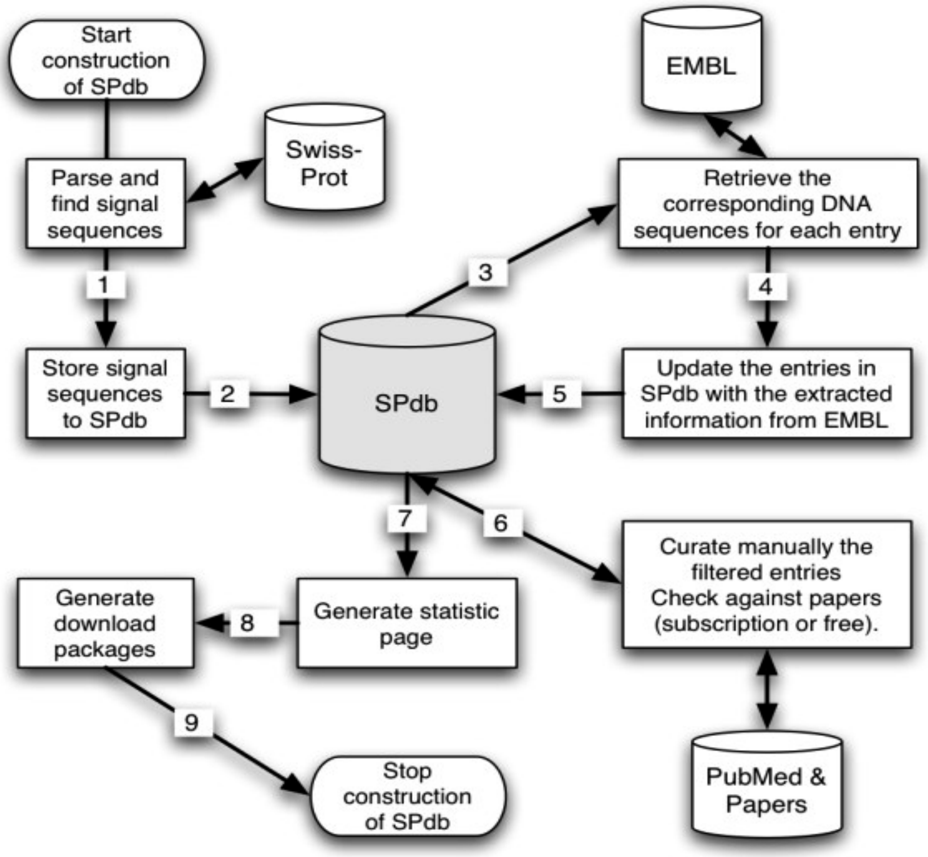
Many researchers face similar obstacles in accessing up-to-date data, which are withheld from public access by method developers (Pennisi 1999; Wiley and Michaels, 2004). Hence, there is an urgent need to provide a publicly accessible, manually curated and regularly updated database specifically for SPs.

3.2 Materials and Methods

To address this critical need, a pipeline (Figure 5) is devised to construct and update a relational database (built using MySQL (http://www.mysql.com/)) to store SP-related information including their sequences. The repository is designed in accordance to the design considerations discussed in the assessment of the available servers and database systems (Tan *et al.*, 2005).

Nielsen *et al.* introduced a methodology (Nielsen *et al.*, 1996) to generate the training and testing datasets for developing the popular SP prediction tool, SignalP (Bendtsen *et al.*, 2004b), which has since undergone two revisions. Some of the proposed criteria are adapted in this work while others are omitted. New criteria are introduced to meet our goal of constructing a high quality repository with accurately annotated SPs.

SPs and coding sequences are extracted from Swiss-Prot (TrEMBL entries are excluded) Entries tagged with the SIGNAL keyword in the feature table FT field (http://www.expasy.org/sprot/userman.html#FT_line) are assumed to contain information on SP. This simple selection process yielded 18,146 entries out of the total 170,140 Swiss-Prot entries (Release 46.1). Entries annotated with keywords such as PROBABLE, POTENTIAL, BY SIMILARITY, HYPOTHETICAL and entries with ambiguous positions (either cleavage sites or starting position) are designated as *unverified sequences*. SPs with length less than 11aa are relegated to the *unverified sequences* set since SPs are generally considered to be of length 15 to 40. Typical SPs with less than 11aa are rarely found in the database. This initial step filtered off 13,701 entries from the *preliminary filtered* set leaving behind 4,445 entries.



Start construction of SPdb
Swiss-Prot
Parse and find signal sequences
1
Store signal sequences to SPdb
2
SPdb
EMBL
Retrieve the corresponding DNA sequences for each entry
3
4
Update the entries in SPdb with the extracted information from EMBL
5
6
7
Curate manually the filtered entries Check against papers (subscription or free).
Generate download packages
8
Generate statistic page
9
Stop construction of SPdb
PubMed & Papers

Figure 5: Schematic diagram of the construction and update protocol of SPdb. The diagram is generated using OmniGraffle (http://www.omnigroup.com).

These entries still include SPs, lipoproteins and Tat-containing signal sequences. Using the SIGNAL keyword, mTP and cTP are indirectly excluded from the *preliminary filtered set* since transit peptides are identified by the TRANSIT keyword in Swiss-Prot.

The next step is to integrate complementary information from EMBL and use that information to check against Swiss-Prot to identify erroneous annotations. This practice of using complementary information from other data sources was found to be useful in data evaluation (Bork, 2000). Here, the first cross-reference to EMBL database is used for each Swiss-Prot entry. Only selected data categories of EMBL are selected (*Appendix B*). Relevant annotations are extracted from EMBL including coding sequence, SP and its length, subcellular location, authors' notes and so on.

The annotations, specifically the *sig_region* and *misc* fields from the EMBL entry are checked against the *preliminary filtered* entries. This step helps to identify many inconsistent entries where the positions have been wrongly quoted by either source, for instance, Swiss-Prot quoted cleavage position of 33 while EMBL provided 32 for the entry (Swiss-Prot ID: CD166_CHICK). Accordingly, another 866 entries are eliminated to yield 3,579 entries in this newly filtered *Swiss-Prot/EMBL* set.

There are some Swiss-Prot entries in the *Swiss-Prot/EMBL* set that are without any EMBL reference e.g. (Swiss-Prot ID: APOE_CAVPO); or lack of annotations in the EMBL entries e.g. (Swiss-Prot ID: 17KD_RICAU); or indicated with annotation such as NOT_ANNOTATED_CDS e.g. (Swiss-Prot ID: 2B31_HUMAN), ALT_TERM e.g. (Swiss-Prot ID: CD1E_HUMAN), ALT_INIT e.g. (Swiss-Prot ID: 1A03_PANTR) and ALT_SEQ e.g. (Swiss-Prot ID: 17KD_RICPR) in their EMBL cross-references. All these entries are earmarked for manual curation. These terms "NOT_ANNOTATED_CDS", "ALT_TERM" and other are known as *status identifiers* and appear in the DR field in Swiss-Prot entries. Detailed explanation is found in the Swiss-Prot manual (http://www.expasy.org/sprot/userman.html#DR_line). The extraction and error detection rules described thus far are collectively known as the '*SP Filtering Rules*' (version 1.0).

Next, with the newly generated filtered set, all the entries in this *Swiss-Prot/EMBL* set are manually checked against the referred publications. Numerous entries with discrepancies in cleavage site between the Swiss-Prot annotations and the accompanying papers are identified e.g. (Swiss-Prot ID: CECC_DROME), the cleavage position was annotated to be 23 in Swiss-Prot while the referenced paper quoted 22. In another entry (Swiss-Prot ID: AMCY_PARVE), Swiss-Prot quoted position 26 while the referenced paper quoted 25 for the cleavage position. For entries which we do not have access to the accompanying papers e.g. (Swiss-Prot IDs: ZEAL_MAIZE, ZEA6_MAIZE) or entries that we are unable to locate their cleavage site information in the papers e.g. (Swiss-Prot ID: GUX1_TRIRE) or entries that are inadequately labeled or those that are specified with inconsistent positional information are all relegated to the *unverified sequences* set.

A further 995 entries are eliminated from the *Swiss-Prot/EMBL* set containing 3,579 entries during the manual curation phase where they are (a) both Swiss-Prot and the quoted papers provided the same putative position (b) different positions were stated by Swiss-Prot and the quoted papers (c) no access to the quoted subscription-only papers or the papers referred to are old and in some cases there were no paper or no relevant paper quoted (d) no mention of cleavage site information (Table 2).

Table 2: A list of the different types of errors that was identified and the problems encountered during the database manual curation step. ^{1} represents the number of entries or sequences identified with the problem described.

Problem description	**No. of entries^1**
Swiss-Prot and the accompanying papers quoted same putative position	311
Swiss-Prot and the accompanying papers quoted different position; The position quoted maybe confirmed or putative	100
No references or relevant references were provided; No access to some subscription-only papers; No access to some very old papers	194
Unable to locate or obtain the position information from papers	390
TOTAL	995

3.3 Results and Discussion

3.3.1 Content of SPdb

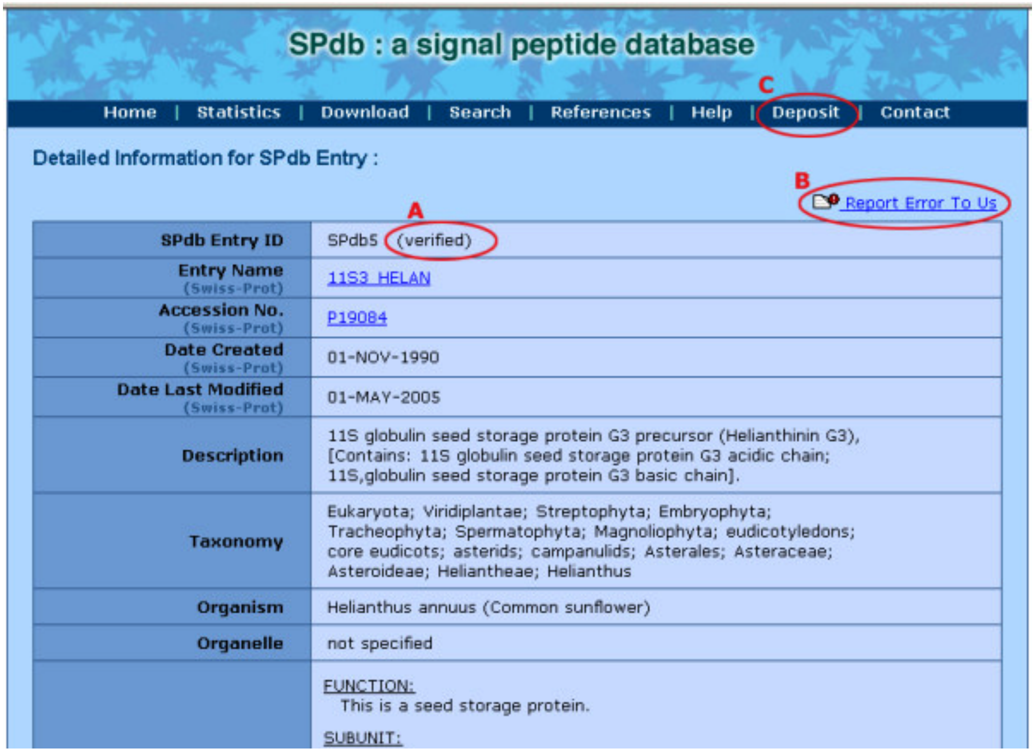
SPdb (release 3.2; http://proline.bic.nus.edu.sg/spdb) was released for public access with a total of 18,146 SP entries in which 2,584 are verified sequences (Table 3). The *verified set* includes lipoproteins, Tat-containing signal sequences and SPs with their mature endogenous proteins that were sequenced on their N-terminal. These entries are manually checked against the accompanying reference paper and they are deemed to contain experimentally verified SPs. The remaining 15,562 *unverified sequences* contain putative or experimentally unverified SP cleavage sites. This *unverified* set potentially contains entries with erroneous annotations and there may be some experimentally verified SPs as well since there are some accompanying papers that we do not have access.

Table 3: Distribution of the sequences organized according to four sub-groups in SPdb 3.2. The verified set in this release of SPdb include SPs, lipoproteins and Tat-containing signal sequences. This practice has been discontinued in subsequent releases of SPdb to include only SPs in the verified set.

	Archaea	Bacteria	Eukaryotes	Viruses	Sub-total
Verified Sequences	7	540	1,945	92	2,584
Unverified Sequences	101	3,528	11,239	694	15,562
TOTAL	108	4,068	13,184	786	18,146

There are 4 data groups in SPdb namely archaea, bacteria, eukaryotes and viruses (Table 3). Information such as organism source, organelle, subcellular location and other accompanying important notes are supplied (Figure 6). Cross-referenced links to the originating database are included as well.

By integrating information from Swiss-Prot and EMBL, SPdb provides a singular point for accessing SPs and the related annotations (Figure 6). An easy-to-navigate web interface written in Perl (Wall, 2000) facilitates user to search through the database with returned results that can be viewed as HTML web page or downloaded in FASTA formated files. User is able to select an entry or a collective set of entries matching the user's criteria such as name of organism, data group, length of signal sequences, keyword searches. Only relevant references that describe the SP are included to allow user to easily consult the corresponding article(s).


SPdb : a signal peptide database
Home | Statistics | Download | Search | References | Help | Deposit | Contact
C
B
A
Report Error To Us

Detailed Information for SPdb Entry :

SPdb Entry ID	SPdb5 (verified)
Entry Name (Swiss-Prot)	11S3_HELAN
Accession No. (Swiss-Prot)	P19084
Date Created (Swiss-Prot)	01-NOV-1990
Date Last Modified (Swiss-Prot)	01-MAY-2005
Description	11S globulin seed storage protein G3 precursor (Helianthinin G3), [Contains: 11S globulin seed storage protein G3 acidic chain; 11S,globulin seed storage protein G3 basic chain].
Taxonomy	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; asterids; campanulids; Asterales; Asteraceae; Asteroideae; Heliantheae; Helianthus
Organism	Helianthus annuus (Common sunflower)
Organelle	not specified
	FUNCTION: This is a seed storage protein. SUBUNIT:

Figure 6: SPdb entry information includes a short description of the protein, the hydropathy plots and amino acids properties and more. (A) Each entry is marked as verified or unverified; (B) An error-feedback link for users to inform us on any error or updated information pertaining to an entry for us to rectify/update; (C) Users can deposit their signal sequences with us and add on their own annotation.

SPdb provides other information such as amino acid composition of the protein which have been suggested to correlate with the subcellular localization of the protein (Nakashima and Nishikawa, 1994); amino acid residues properties (aromatic, non-polar, polar, charged and so on) are shown in graphical format to indicate which residues possess the properties visibly; also accompanying each entry are various the hydropathy plots (Kyte and Doolittle, 1982; Sweet and Eisenberg, 1983; Eisenberg *et al.*, 1982) of the SPs and the sequences downstream of the SP cleavage site. The plots are rendered using *pepinfo* found within the computational analysis package of EMBOSS (Rice *et al.*, 2000), an open source software suite for sequence analysis. Each SP exhibits three distinct regions at the sequence level: the *n-region* (a positive charged region), the *h-region* (hydrophobic region) and the *c-region* (polar and neutral region) (von Heijne, 1990). The hydropathy plots thus help in visualizing these regions for easy identification purpose.

It was shown that SP processing by the SRP requires certain contextual cues in the sequence downstream. SRP binds to n-terminus signal or signal-anchor sequence when the nascent polypeptide chain is synthesized by the ribosome up to approximately 60aa. At this length, this segment is conveniently exposed and translation will resume upon dissociation of SRP from the nascent chain (de Gier *et al.*, 1998). In the effort to capture this information for the co-translational translocation mechanism, SPdb includes both the SPs and 30aa after the cleavage site and they are colored using the RasMol amino acids color scheme (http://www.openrasmol.org/doc/rasmol.html#aminocolours) with explicit mark of the SP cleavage site.

3.3.2 Experimental support in database entries

Swiss-Prot, a venerable and often cited gold standard for manually curated database, has been the authoritative source for more reliable sequence entry annotation. Often, it is presumed that an entry should be relatively accurate particularly if there is no label that indicates it is lacking in experimental support. As a result, the data are usually extracted without much further processing or examination. However, as we have shown in this study, multiple types of erroneous exist. For instance, conflicting annotations are reported on the positions or length of the signal sequences by Swiss-Prot when compared against EMBL e.g. (Swiss-Prot IDs: A2AP_HUMAN, BTD_HUMAN). Inconsistencies such as this usually arise when there is more than one reference. The referred papers may quote different positions that may cause the confusion. By combining annotations from EMBL and using it to crosscheck against Swiss-Prot, we have managed to identify many such entries. The annotations on SP found in EMBL are relatively more accurate though there are incidents when the information is incorrect e.g. (EMBL ID: M19077 for the entry Swiss-Prot ID: CHR1_BOMMO). To tackle this, a link is provided in SPdb to allow the user to report any error or discrepancy that is encountered in SPdb.

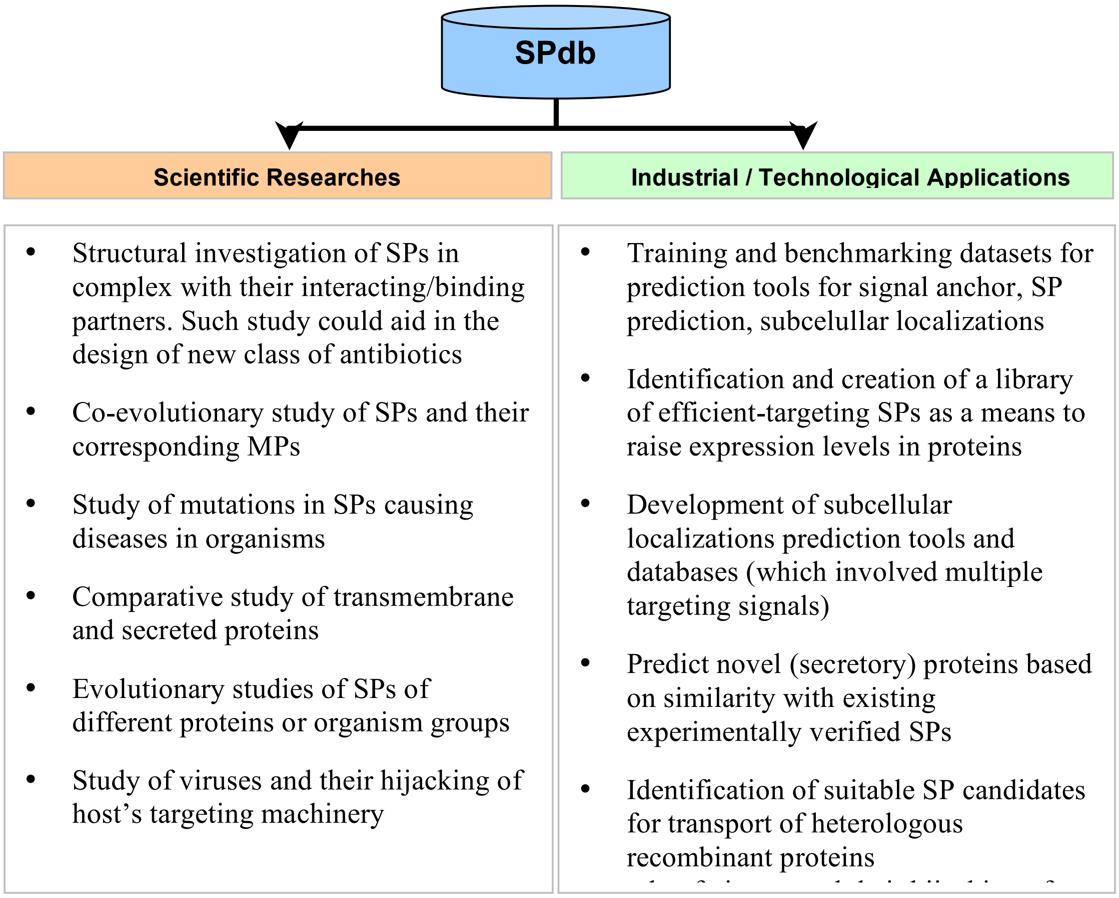
There is also the issue on experimental evidence support provided in the journal publication. Many of the entries with annotations found in the sequence databases on SP position are predicted or deemed putative or potential as reported by the researchers in the accompanied literature (Table 2). Unfortunately, these entries are not properly labeled with keywords like POTENTIAL, BY SIMILARITY and PROBABLE as previously assumed in Swiss-Prot. Many of the referred papers actually used prediction tools or sequence alignment software to identify the possible SP cleavage site. This has serious implication on many downstream works which rely on Swiss-Prot and other primary databases. Often, they assume that such entries are experimentally verified since there was no indication that suggested otherwise.

3.3.3 Text-mining as an extraction method

Prior to our manual curation effort, we have explored text-mining approach as the technique has been applied in the extraction of protein-protein interactions (Thomas *et al.*, 2000), protein structure and residues (Gaizauskas *et al.*, 2003), full-text biomedical article (Corney *et al.*, 2004), gene/protein biological function (Koike *et al.*, 2005), albeit with moderate success. However, we soon discovered that many of abstracts do not contain cleavage site information. The information is often located in the body of the paper, usually appearing under the results or discussion section. In some cases, the cleavage site is embedded within an image file either indicated with an arrow or asterisk. Further complicating the matter are the words or phrases that were used to express the positional information, for example, in the paper (Hinuma *et al.*, 1998) quoted in entry (Swiss-Prot ID: PRRP_BOVIN), there is the sentence "… its N-terminal portion before Ser-23 showed the typical profile of a secretory signal peptide …". Such sentences often vary and it is difficult to extract such information by using extraction rules. Additionally, many of the papers are view-by-subscription only, rendering the extraction program useless. Unless each of the paper submitted in future provides a short note on the features of the proteins described coupled with the improvement in text-mining accuracy, it will be extremely difficult and text-mining approach can only be applied sparingly. Manual curation will still be required for the time being although open-access journals are increasingly prevalent (http://www.doaj.org/). The availability of full-text articles may enable wider adoption of text-mining tools to extract information from the articles.

3.3.4 Uses of SPdb

Although SPdb was mainly created to support the studies in this work, the curated SPdb can be a valuable and useful resource to support further scientific studies into multiple areas. It is also applicable to technological or industrial domains. Figure 7 briefly describes some potential uses of SPdb.



SPdb

Scientific Researches

- Structural investigation of SPs in complex with their interacting/binding partners. Such study could aid in the design of new class of antibiotics
- Co-evolutionary study of SPs and their corresponding MPs
- Study of mutations in SPs causing diseases in organisms
- Comparative study of transmembrane and secreted proteins
- Evolutionary studies of SPs of different proteins or organism groups
- Study of viruses and their hijacking of host's targeting machinery

Industrial / Technological Applications

- Training and benchmarking datasets for prediction tools for signal anchor, SP prediction, subcelullar localizations
- Identification and creation of a library of efficient-targeting SPs as a means to raise expression levels in proteins
- Development of subcellular localizations prediction tools and databases (which involved multiple targeting signals)
- Predict novel (secretory) proteins based on similarity with existing experimentally verified SPs
- Identification of suitable SP candidates for transport of heterologous recombinant proteins

Figure 7: Potential uses of SPdb in scientific researches and technological applications.

For the purpose of this work, we shall exploit SPdb to analyze the SP sequences to understand the factors that contribute to their SP cleavage processing. It will also be used for structural study of SPs to further our understanding of how the structural conformations contribute to the substrate specificity. These works shall be discussed in the following chapters.

3.4 Summary

A semi-automated pipeline driven by our "*SP Filtering Rules*" (version 1.0) has been developed for the generation of SPdb, a repository that is dedicated for the study of SPs. The resulting entries are further curated manually to ensure that experimentally verified SPs are identified and sufficiently annotated.

New error detection techniques were devised and the integration of information from different databases has helped to eliminate inconsistent and erroneous entries. The provision of this system drastically reduces the laborious effort required to assemble SP related data on a regular basis and at the same time, it ensures generation of consistent and standardized datasets. SPdb can be a useful resource for prediction and analysis works, and it provides the much-needed standardized dataset for benchmarking the SP prediction tools. SPdb can also serve as the foundation for exploration into other scientific studies as well as supporting the development of technological applications.

For future releases, there are plans for further classifications, for instance, according to functions, conservations of SP sequences and so on. As differentially targeted organelles or locations are variations on the general theme of SP targeted proteins, it would be logical to include these different targeting signals for comparison and studies purpose. Further, information on secreted proteins that lack cleavable SP (Ye *et al.*, 1988) e.g. ovalbumin, a secreted glycoprotein and the major protein of egg-white which does not have a cleavable SP (Lingappa *et al.*, 1979; Belin *et al.*, 2004) could be included. This can allow us to compare the differences between cleavable and non-cleavable SPs.

Chapter 4: Sequence Analysis of SPs

4.1 Introduction

When the "*signal hypothesis*" (Blobel and Dobberstein, 1975a; Blobel and Dobberstein, 1975b) was first mooted in 1970s, investigation into SPs was still fairly limited. The pace has since hastened, with various studies scrutinizing different aspects of these transient polypeptides. Large-scale efforts such as the SPDI (Clark *et al.*, 2003) were launched to identify novel secreted and transmembrane proteins in human while other project such as the Human Proteome Folding Phase II (http://www.worldcommunitygrid.org/projects_showcase/viewHpf2About.do) focuses on predictive, preventative and personalized medicine where human secreted proteins and pathogenic proteins are clearly form the key elements. Such initiatives and the on-going sequencing work at the research labs and sequencing centers around the globe continue to generate large number of sequences and new insights.

The deluge of protein sequences has spurred the active development of computational tools and techniques to automate the prediction of SPs (*Chapter 6*: Table 5). While the accuracies of these predictive tools vary depending on the datasets employed in their studies, they have generally achieved good accuracy (80-90%). Nonetheless, the precise mechanism governing the cleavage of the preprotein thus far remains a conundrum. The accuracy of even the best prediction methods for mutations/alterations to the SP region remains unpredictable. In order to understand the SP cleavage processing and targeting mechanism, it is necessary to first examine the SP and MP sequences.

An early study involving 118 eukaryotic and 32 prokaryotic sequences provided an excellent insight into the nuances of eukaryotic and bacterial SPs (von

Heijne, 1985). This was later followed by a study of 900 eukaryotic and 200 prokaryotic sequences with known cleavage sites (von Heijne and Abrahmsén, 1989). Subsequent studies (McKnight *et al.*, 1991; Rajalahti *et al.*, 2007; Thornton *et al.*, 2006) investigated SPs and MPs, either singularly or in combination, often through gene fusion and mutagenesis studies to observe their translocation and differential expression levels. Wide range of studies (Biro, 2006; Eusebio *et al.*, 1998; Kajava *et al.*, 2000; Kantardjieff and Rupp, 2004; Matoba and Ogrydziak, 1998; Tsuchiya *et al.*, 2003; von Heijne, 1986b) were conducted to inspect the charge bias and hydrophobicity of SPs. Austen *et al.* employed synthetic SPs to demonstrate that it is the common structural features incorporated that bestowed the SPs their translocation function, and not simply the primary sequence (Austen *et al.*, 1984). Structural studies investigated SPase I-substrate complexes through 3D-structures and computational models (Ekici *et al.*, 2007; Paetzel *et al.*, 2000; Paetzel *et al.*, 2002a) to study the substrate specificity and the characteristics of the amino acid residues around the cleavage site. With the massive increase in deposition of protein sequences into the public sequence databases since the last sequence study involving larger dataset with known cleavage sites (Nielsen *et al.*, 1997), there is a tremendous opportunity to improve our understanding of SPs.

In this study, 2,352 eukaryotic and bacterial SPs are extracted from SPdb (Choo *et al.*, 2005) following an improved protocol from the original (*Chapter 3*). The aim is to examine the characteristics of the amino acid residues at the cleavage site. Furthermore, the residue composition in the vicinity of the cleavage site are investigated, as a multitude of site-directed mutagenesis studies have revealed that residues upstream and downstream of this site affect cleavage site recognition and processing (Kajava *et al.*, 2000; Russel and Model, 1981).

4.2 Materials and Methods

4.2.1 Data preparation using SPdb

A preliminary dataset of 2,512 sequences is assembled from the manually curated SPdb 5.1. Only sequences reported with experimentally verified SP cleavage sites are used. CD-HIT (version 3.1.1) (Li and Godzik, 2006) is used to cluster and remove sequences with 100% sequence identity in their SP moiety as studies (Jarjanazi *et al.*, 2008; Rajpar *et al.*, 2002; Tsujibo *et al.*, 1994) have shown that even a single substitution of the amino acid in SP could result in a pronounced effect.

Next, the dataset is split into two sub-datasets based on their (i) SP and (ii) MP before being clustered again using CD-HIT, with global sequence identity threshold set at 0.9 (90%); word size of 5 and other parameters set at the program's default. The reduced dataset of 2,352 SP-containing sequences is further categorized into five groups: (i) Gram+ bacteria (*Firmicutes, Actinobacteria, Deinococcus, Fibrobacteres, Thermotogae*) *Mollicutes* which are lack of cell wall are excluded; (ii) Gram- (*Proteobacteria, Spirochetes, Bacteroidetes, Cyanobacteria, Aquificae, Chlamydiae*) and (iii) eukaryotes (iv) archaea (v) viruses. Viral and archaeal SPs are retained for analysis in subsequent study, as there are only a few sequences with experimental support. The bacteria dataset is classified into Gram+ and Gram- due to their distinctive features (von Heijne, 1990). The physico-chemical properties of the SP and MP for each sequence are computed using ExPASy ProtPram (Walker, 2005). Other calculations include molecular weight, theoretical isoelectric point (p*I*), aliphatic index, GRand AVerage of hydropathY (GRAVY) and absolute mean charge.

4.2.2 Calculations of the physico-chemical properties

Size dimension is assumed to influence the bending of a peptide chain where the size of an amino acid is determined by the length and bulkiness of its side chain (Biro, 2006). Since molecular weight of an amino acid is easier to measure and it is roughly proportional to its size, hence, we use it as an approximation.

p*I* is defined as the pH value where a given protein has no net charge and it often has the lowest solubility. Different algorithms exist to calculate p*I* rendering different values due to the different set of pK_a values used. The pK_a values adopted in this study were described (Bjellqvist *et al.*, 1993).

Aliphatic index (Ikai, 1980) measures the relative volume occupied by aliphatic side chains (Ala, Val, Ile and Leu) of a protein according to the formula:

$$AliphaticIndex = X_A + a * X_V + b * (X_I + X_L)$$

where X_A (Ala), X_V (Val), X_I (Ile) and X_L (Leu) are mole percent (100 * mole fraction) of the respective amino acid residue. The coefficients *a* and *b* represent the relative volume of Val side chain (*a* = 2.9) and of Leu/Ile side chains (*b* = 3.9) compared to the side chain of Ala.

GRAVY (Kyte and Doolittle, 1982) is an estimation of the overall hydrophobicity of a protein, but it does not consider interaction or positional effect of adjacent residues. Given a protein sequence *S*, its GRAVY score is computed as:

$$GRAVY(S) = \sum_{i=1}^{20} \alpha_i f_i$$

where *i* is one of the twenty standard amino acids; f_i is the relative frequency of *i* in *S*; α_i is the hydropathy value of *i* according to the scale propounded by (Kyte and Doolittle, 1982) and *n* is the total number of residues in the sequence.

Net charge is the algebraic sum of all the charged amino acid residues present in SPs and MPs calculated using the equation:

$$\text{Net Charge} = \sum_{i=1}^{20} \alpha_i f_i$$

The twenty standard amino acids are represented by *i* and f_i represents the relative frequencies of occurrences of the amino acid *i*. Positively-charged residues (Arg, His and Lys) are assigned $\alpha_i = 1$ whereas negatively-charged residues (Asp and Glu) are set as $\alpha_i = -1$. All other amino acid residues are assigned $\alpha_i.= 0$.

The iep program, from the EMBOSS bioinformatics package (version 2.9.0) (Rice *et al.*, 2000) was used to calculate the mean charge at neutral pH. The absolute value of the mean charge is further divided by the length of the polypeptide.

Mean hydrophobicity is defined as the arithmetic mean of the normalized hydrophobicity values of all the residues in the polypeptide (Kyte and Doolittle, 1982).

4.3 Results

4.3.1 Datasets

A curated set of 2,352 SP-containing sequences is assembled for this study using SPdb 5.1 (http://proline.bic.nus.edu.sg/spdb/analysis.html).

Scatter plots of the assembled SPs for the three groups are generated and β-hexosaminidase A (Swiss-Prot ID: HEXA_PSEO7), an αβ-subunit heterodimer lysosomal hydrolase was identified as an outlier. Tsujibo and co-workers indicated that the SP cleavage site is 11aa and added that the SP does not possess the typical tripartite features of an SP (Tsujibo *et al.*, 1994). However, sequence comparison against other species using Swiss-Prot database reveals lengths of approximately 18 to 22aa. Due to this inconsistency, this entry was removed from the final dataset.

4.3.2 Examining the eukaryotic and bacterial datasets

The cleansed dataset is grouped into (i) eukaryotes (Euk) with 1,877 sequences (ii) Gram+ bacteria with 168 sequences and (iii) Gram- bacteria with 307 sequences.

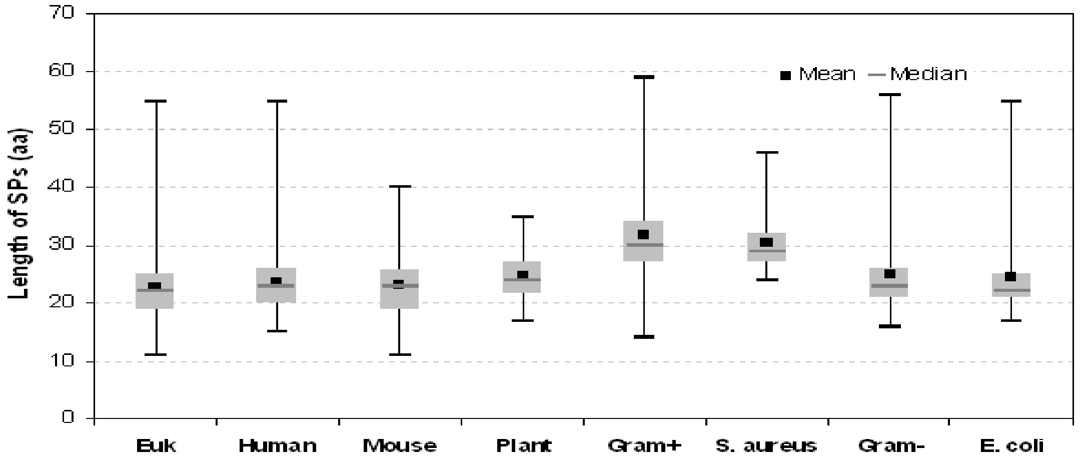


Figure 8: Boxplot illustrating the SPs distribution found in selected organisms and groups (eukaryotes, Gram+ and Gram- bacteria). Mean length (■) and median (—, gray bar) values are indicated.

The boxplot (Figure 8) shows the length distribution of SPs for the different organism groups. Sub-groups from these organism groups are plotted as well. The distinctive long whiskers as seen in all the boxes affirm the previous studies that have reported SPs as having variable length. From the boxplot, SPs of Gram+ (SPs_{Gram+}) are clearly longer with median length of 30aa and display a bi-modal distribution with peaks at 29aa and 41aa (Figure 9), compared to SPs of eukaryotes (SPs_{Euk}) and SPs of Gram- (SPs_{Gram-}) bacteria which carry median length of 22aa and 23aa respectively. Interestingly, SPs_{Euk} and SPs_{Gram-} exhibit somewhat similar SP length distribution although 4.5% or fourteen SPs_{Gram-} extends beyond 40aa. Despite the wide range of SP lengths permissible within many groups of organisms excluding SPs of plants (SPs_{Plant}), majority lengths appear to fall in the 25th to 75th percentile.

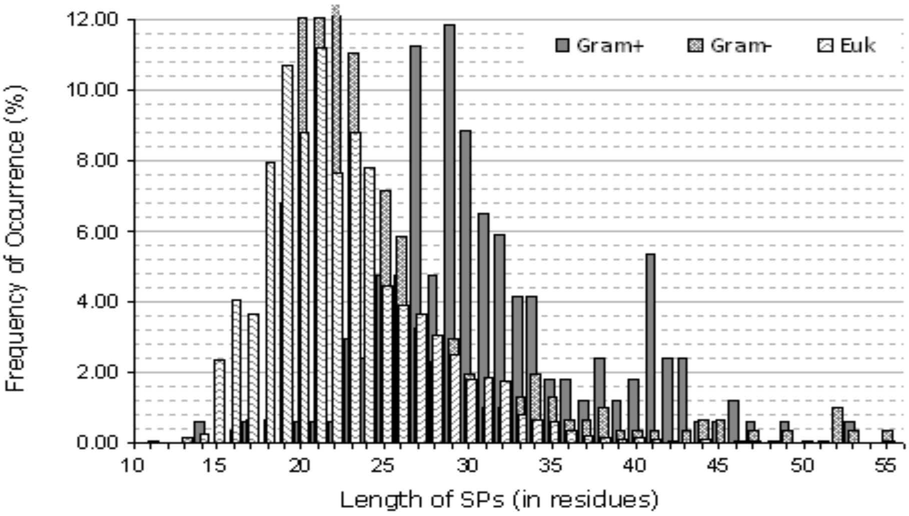


Figure 9: SPs from the three organism groups measured based on their length. The Y-axis shows the frequency of occurrences for a specific length of SP while the X-axis depicts the various lengths.

Next, we examine the occurrences of amino acid residues at different positions. Figure 10 depicts the sequence logos (Crooks *et al.*, 2004) for the three groups starting from position P35 to position P5', spanning contiguous segments from the SP and MP. The cleavage site occurs between P1 and P1'.

P1 and P3 favor small, aliphatic residues; in particular Ala and Val, with striking inclination that is apparent in bacterial SPs. Gly, Ser and Thr are also noticeable at these two positions in SPs_{Euk}. P2 of SPs_{Euk} exhibits preferences for Leu (15.2%) and Ser (12.0%) whereas different sets of amino acids: [Ser (12.5%), Gln (11.9%), Phe (11.9%), Ala (11.3%)] and [Leu (17.6%), Gln (14.3%), Phe (11.4%), His (11.4%)] are preferred by SPs_{Gram+} and SPs_{Gram-} respectively (Table 4). From P1' onwards, there is no obvious pattern of amino acid conservation in SPs_{Euk} with the exception of slightly enhanced occurrences of Ala (13.5%) and Gln (11.0%) at P1'.

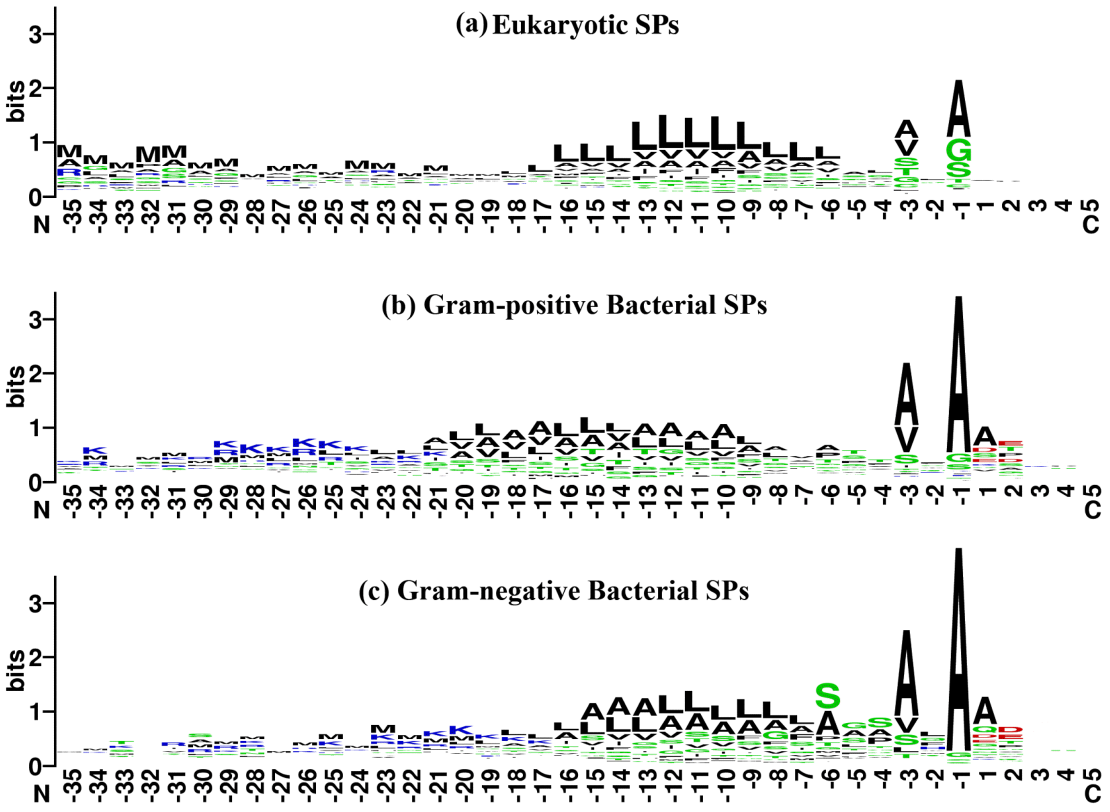


Figure 10: Sequence logos (Crooks *et al.*, 2004) of eukaryotic and bacterial (Gram+ and Gram-) SPs and MPs starting from P35 to P5'. The interface between P1 and P1' represents the SPase I cleavage site. The amino acid residues are grouped and colored based on the R group of their side chain. Red denotes polar acidic amino acid residues (D,E); Blue denotes polar basic amino acid residues (K, R, H); Green denotes polar uncharged amino acid residues (C, G, N, Q, S, T, Y); Black denotes non-polar hydrophobic amino acid residues (A, F, I, L, M, P, V, W).

Compared to SPs$_{\text{Euk}}$, the amino acid composition is different in bacterial SPs. In SPs$_{\text{Gram+}}$, P1' is mostly occupied by Ala (36.3%), Asp (11.3%), Ser (10.7%) and Glu (9.5%). P2' is populated by Thr (14.3%), Glu (13.7%), Pro (13.1%), Ser (10.7%) and Asp (10.7%). Lys (13.1%) is the dominant amino acid at P3' while Pro (14.3%) and Thr (14.3%) are preferred at P4'. Beyond P4', there is no clear pattern if the relative frequencies are compared between the adjacent positions for the same amino acid type. Similarly for SPs$_{\text{Gram-}}$, P1' is populated by Ala (41.7%), Gln (12.1%), Asp (7.2%) and Glu (6.2%) whereas P2' is largely distributed between Asp (17.3%), Glu (16.9%), Pro (10.8%) and Thr (10.8%). From P3' onwards, when the relative frequencies of each amino acid are compared to its adjacent positions and also within the column (Table 4 and Figure 10), there is no discernible pattern. His, Trp and Tyr are clearly under-represented in all three groups of SPs and for all the positions (P10 to P10') that are examined while Cys is almost nonexistent in bacterial SPs throughout the mentioned positions. Pro is visibly avoided in positions from P3 to P1' but relatively prevalent at P4 and P2'. In contrast, Gly, Ile, Thr (except at P1 in bacterial SPs), Val (except at P1), Ser and particularly Ala (especially at P3, P1 and P1') are ubiquitous in all the positions that we profiled.

The occurrence of acidic residues (Asp and Glu) is pronounced from P1' onwards in all three groups of SPs. Similar trends can be seen for basic or positively charged residues comprising Arg, Lys and His. In fact, when the basic and acidic residues are grouped (Table 4), there is a consistent and modest occurrence of these charged residues across all three groups of SPs from P1' onwards, inclusive of P2 but conspicuously absent or appearing in minute amounts at P3 and P1, most prominently in the eukaryotic MPs. Basic residues, Arg and Lys are common at the n-region of bacterial SPs.

Table 4: Amino acid frequency matrix for the SPs and MPs of eukaryotes and bacteria. Percentage occupancy values from P10 to P10' [+10, -10] are shown, with the cleavage site represented by dotted line at the -1/+1 junction. Significant high and low values are highlighted: gray: >10%; black: most preferred residue(s); cyan: charged residue group and green: aliphatic group.

Eukaryote	**-10**	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**	**+4**	**+5**	**+6**	**+7**	**+8**	**+9**	**+10**
Ala(A)	9.43	16.25	13.43	11.51	13.11	15.18	9.38	**25.84**	7.46	**48.91**	13.53	3.68	4.74	5.33	5.17	5.49	5.01	5.27	5.43	6.23
Cys(C)	3.68	4.05	5.91	5.81	3.68	2.18	3.36	6.50	1.49	3.57	3.41	1.92	7.73	4.16	3.46	5.01	4.85	3.84	3.84	3.89
Asp(D)	0.11	0.16	0.27	0.43	0.69	3.84	1.33	0.27	3.94	0.32	5.33	7.99	5.49	5.59	5.65	5.65	5.97	5.75	4.90	5.59
Glu(E)	0.32	0.32	0.85	1.39	0.69	2.02	2.66	0.37	8.79	0.53	7.83	8.84	5.22	7.19	7.25	6.45	6.02	6.71	5.97	6.07
Phe(F)	9.11	6.87	5.86	6.77	5.91	1.60	2.34	0.64	3.46	0.21	3.57	2.18	3.62	3.36	3.14	3.94	3.09	4.53	3.04	3.94
Gly(G)	4.00	3.94	5.27	4.32	3.46	11.13	13.96	9.43	3.52	**20.72**	7.14	5.01	6.29	8.10	6.66	7.19	5.91	8.15	8.95	8.58
His(H)	0.27	0.48	0.80	0.75	0.59	2.50	1.23	0.16	4.95	0.05	1.81	2.45	2.45	2.61	3.09	2.18	3.62	1.81	2.29	2.40
Ile(I)	5.17	6.18	4.37	4.05	7.94	2.29	3.41	3.78	1.70	0.16	3.30	3.73	6.23	2.72	3.14	3.46	3.14	3.57	4.79	3.52
Lys(K)	0.11	0.00	0.05	0.91	0.16	2.08	1.92	0.37	1.44	0.11	4.64	4.95	2.45	4.79	5.38	4.85	4.85	3.84	5.27	5.22
Leu(L)	**43.79**	**37.93**	**29.89**	**36.49**	**27.22**	7.94	16.41	4.32	15.24	1.39	8.47	4.58	9.11	5.65	5.97	7.46	6.55	7.67	8.26	7.25
Met(M)	2.40	1.97	3.30	2.50	1.81	1.17	1.70	0.27	1.76	0.21	1.12	0.69	1.70	1.28	1.70	1.17	1.33	2.13	1.44	1.70
Asn(N)	0.69	0.48	0.59	1.23	0.75	2.02	0.91	0.69	4.21	0.37	2.50	4.32	3.73	3.94	4.32	3.25	5.43	4.37	4.26	4.95
Pro(P)	1.01	1.17	0.96	2.34	6.29	9.38	9.11	0.21	0.69	2.02	0.27	15.82	7.46	10.66	8.79	8.58	8.68	9.86	6.93	6.55
Gln(Q)	0.64	0.75	1.97	1.65	1.07	5.27	4.48	0.27	7.25	1.33	11.03	4.90	3.52	6.77	5.54	5.54	5.06	3.84	6.93	4.48
Arg(R)	0.11	0.21	0.37	1.17	0.75	2.98	2.88	0.37	5.54	0.96	4.95	4.69	2.29	3.62	5.59	6.13	4.58	5.49	4.00	4.21
Ser(S)	3.30	6.23	7.88	6.29	6.93	12.20	8.47	13.00	11.99	13.48	8.20	9.00	7.62	7.94	7.94	6.55	6.87	8.15	8.31	6.61
Thr(T)	4.26	3.20	4.53	3.62	4.16	8.90	6.39	10.97	4.05	5.01	4.26	5.59	6.61	7.51	7.46	6.39	6.82	5.59	4.90	7.51
Val(V)	10.28	8.74	10.87	5.81	12.36	5.17	7.46	**22.32**	3.57	0.37	4.48	6.34	8.74	5.11	5.70	7.62	6.39	5.65	6.55	6.39
Trp(W)	0.91	0.64	1.76	2.13	1.65	1.17	1.23	0.05	4.21	0.16	1.01	0.85	1.86	0.80	0.69	0.75	1.07	1.65	1.01	1.76
Tyr(Y)	0.43	0.43	1.07	0.85	0.80	0.96	1.39	0.16	4.74	0.11	3.14	2.45	3.14	2.88	3.36	2.34	4.74	2.13	2.93	3.14
Charged	**0.91**	**1.17**	**2.34**	**4.64**	**2.88**	13.43	10.02	**1.55**	24.67	**1.97**	24.56	28.93	17.90	23.81	26.96	25.25	25.04	23.60	22.43	23.49
Small	16.73	26.43	26.58	22.11	23.49	38.52	31.81	48.27	22.96	**83.11**	28.88	17.69	18.65	21.36	19.77	19.23	17.79	21.58	22.70	21.42
Aliphatic	**68.67**	**69.10**	**58.55**	**57.86**	**60.63**	30.58	36.65	**56.26**	27.97	**50.83**	29.78	18.33	28.82	18.81	19.98	24.03	21.10	22.16	25.04	23.39
Hydro-phobic	38.41	41.82	40.60	36.01	49.23	38.04	36.55	**53.49**	24.29	**52.16**	31.91	38.25	36.81	34.04	33.72	35.86	33.56	36.49	34.47	35.32
Polar uncharged	17.00	19.07	27.22	23.76	20.83	42.67	38.95	41.02	37.24	44.59	39.69	33.19	38.63	41.29	38.73	36.28	39.69	36.07	40.12	39.16

Gram +ve	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
Ala(A)	**25.60**	16.67	14.88	11.90	19.05	7.14	10.12	**51.79**	11.31	**83.93**	**36.31**	9.52	7.14	6.55	11.90	7.74	10.71	10.71	10.12	5.95
Cys(C)	0.60	1.19	0.60	0.60	0.00	0.00	0.00	0.00	0.60	0.00	0.60	0.00	0.60	1.19	0.60	0.60	0.00	0.60	0.60	0.00
Asp(D)	0.00	0.60	0.60	1.19	2.38	2.98	4.17	0.00	2.38	0.00	11.31	10.71	7.14	6.55	9.52	7.14	8.33	7.14	8.33	7.14
Glu(E)	1.79	0.00	0.60	1.79	1.19	3.57	4.76	1.79	5.36	0.00	9.52	13.69	5.36	4.17	8.33	5.95	5.95	6.55	5.95	8.33
Phe(F)	7.74	9.52	4.76	7.74	1.79	1.19	0.60	1.19	11.90	0.00	2.38	1.19	3.57	2.38	3.57	4.17	1.79	0.00	1.79	2.98
Gly(G)	6.55	5.95	11.90	7.14	7.14	8.33	6.55	1.79	4.17	4.76	1.19	7.74	8.93	8.33	5.36	6.55	4.17	4.76	6.55	7.14
His(H)	0.00	0.00	0.60	1.19	1.79	0.60	1.19	0.00	6.55	0.00	1.19	1.79	0.60	1.19	1.19	0.60	1.19	1.19	1.79	1.19
Ile(I)	5.36	6.55	5.36	5.36	4.76	1.79	3.57	2.98	1.19	0.00	1.79	1.19	3.57	3.57	2.38	4.17	7.14	2.98	2.98	2.98
Lys(K)	0.00	0.00	0.60	0.60	2.38	4.17	5.36	0.60	8.93	2.38	5.36	1.19	13.10	6.55	5.95	5.36	7.14	7.14	8.93	7.14
Leu(L)	19.64	**20.83**	15.48	11.90	5.36	3.57	4.76	1.19	5.36	1.19	1.19	1.19	5.36	2.98	5.36	4.17	4.76	6.55	8.33	4.76
Met(M)	1.19	2.98	5.95	3.57	1.19	0.60	3.57	0.60	1.79	0.00	0.60	0.00	0.60	0.60	0.60	0.60	1.79	1.79	1.79	1.19
Asn(N)	2.38	2.98	1.79	4.76	1.79	7.74	5.36	0.00	4.76	0.00	2.98	4.17	5.95	7.14	9.52	5.95	4.17	8.93	5.36	6.55
Pro(P)	0.00	5.95	4.76	5.95	17.26	11.90	12.50	0.00	0.60	0.60	0.00	13.10	7.14	14.29	3.57	9.52	7.74	7.14	7.14	2.38
Gln(Q)	0.60	0.60	1.19	5.36	5.36	5.95	3.57	0.60	11.90	0.00	5.36	1.79	4.17	4.17	4.17	7.74	3.57	2.98	3.57	5.95
Arg(R)	1.19	0.00	0.00	0.00	0.60	1.19	1.19	1.79	2.38	1.19	1.79	0.60	2.38	1.19	1.19	1.19	2.38	1.79	1.19	4.17
Ser(S)	6.55	10.71	10.12	7.14	7.74	16.07	6.55	9.52	12.50	3.57	10.71	10.71	8.33	5.36	7.74	4.17	8.93	8.93	8.93	8.33
Thr(T)	7.74	4.76	10.71	10.71	13.10	16.07	16.67	2.38	1.79	1.19	2.38	14.29	7.74	14.29	8.93	10.12	11.31	8.93	7.74	9.52
Val(V)	11.90	9.52	7.74	11.90	6.55	6.55	8.33	**23.81**	3.57	1.19	4.17	5.95	6.55	7.14	3.57	9.52	6.55	7.74	5.95	9.52
Trp(W)	0.60	1.19	1.19	1.19	0.60	0.60	0.60	0.00	0.60	0.00	0.60	0.00	0.00	0.60	1.19	1.79	1.19	0.00	0.60	1.79
Tyr(Y)	0.60	0.00	1.19	0.00	0.00	0.00	0.60	0.00	2.38	0.00	0.60	1.19	1.79	1.79	5.36	2.98	1.19	4.17	2.38	2.98
Charged	**2.98**	**0.60**	**2.38**	**4.76**	**8.33**	12.50	16.67	**4.17**	25.60	**3.57**	29.17	27.98	28.57	19.64	26.19	20.24	25.00	23.81	26.19	27.98
Small	38.69	33.33	36.90	26.19	33.93	31.55	23.21	**63.10**	27.98	**92.26**	48.21	27.98	24.40	20.24	25.00	18.45	23.81	24.40	25.60	21.43
Aliphatic	**62.50**	**53.57**	43.45	41.07	35.71	19.05	26.79	**79.76**	21.43	**86.31**	43.45	17.86	22.62	20.24	23.21	25.60	29.17	27.98	27.38	23.21
Hydro-phobic	**52.38**	**52.38**	45.24	48.21	**53.57**	33.93	44.64	**80.95**	39.88	**88.10**	**51.19**	32.14	41.67	41.67	32.74	42.86	44.05	37.50	39.29	33.93
Polar uncharged	25.00	26.19	37.50	35.71	35.12	**54.17**	39.29	14.29	38.10	**9.52**	23.81	39.88	37.50	42.26	41.67	38.10	33.33	39.29	35.12	40.48

Gram -ve	**-10**	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**+1**	**+2**	**+3**	**+4**	**+5**	**+6**	**+7**	**+8**	**+9**	**+10**
Ala(A)	**22.48**	**23.45**	16.94	16.94	**30.62**	16.94	16.61	**61.89**	6.19	**93.16**	**41.69**	5.86	8.14	8.14	11.40	8.79	9.45	8.47	7.49	11.40
Cys(C)	1.95	1.95	3.58	0.65	1.30	0.98	0.65	0.65	0.00	0.00	0.00	0.00	1.30	0.00	0.65	0.65	0.00	0.98	0.65	1.95
Asp(D)	0.33	0.33	0.33	0.00	0.00	0.00	0.33	0.00	0.00	0.00	7.17	17.26	3.58	9.45	5.21	6.19	6.51	5.21	4.56	6.84
Glu(E)	0.65	0.00	0.65	0.00	0.65	0.65	0.33	0.00	0.65	0.00	6.19	16.94	4.56	5.86	8.14	8.14	6.51	5.54	6.19	5.54
Phe(F)	4.56	7.49	7.17	14.33	1.95	7.17	2.93	0.98	11.40	0.00	0.98	0.98	3.58	1.30	1.95	1.95	4.89	3.26	1.95	2.61
Gly(G)	6.51	3.91	14.98	6.84	4.23	17.26	5.21	0.98	0.98	2.93	6.19	6.51	7.49	6.51	5.54	8.14	6.51	12.38	10.42	6.84
His(H)	0.00	0.00	0.33	0.65	0.00	1.63	0.65	0.33	11.40	0.33	0.98	0.33	0.98	0.98	1.63	2.61	1.95	0.98	1.30	0.65
Ile(I)	4.89	4.89	2.28	7.17	2.61	0.33	1.95	1.30	1.95	0.00	0.33	2.93	4.56	5.21	6.51	4.56	6.51	6.19	6.19	5.21
Lys(K)	0.00	0.33	0.00	0.33	0.00	0.33	0.65	0.00	0.98	0.00	2.93	1.63	6.19	3.58	2.93	8.14	4.23	7.17	6.51	10.75
Leu(L)	**26.71**	**30.29**	**28.01**	19.54	2.93	9.45	3.58	5.86	17.59	0.33	3.26	1.63	8.14	7.82	6.84	5.21	4.89	7.17	5.54	4.89
Met(M)	3.26	3.58	3.58	5.21	3.91	3.91	1.30	0.00	6.19	0.00	0.00	0.00	1.30	0.98	1.30	0.98	3.26	0.98	1.30	1.63
Asn(N)	0.65	0.65	0.00	0.98	0.98	3.26	4.56	0.33	4.56	0.33	3.58	5.86	5.21	6.19	9.45	6.19	4.23	3.91	8.47	4.89
Pro(P)	0.98	0.98	0.33	1.30	6.51	5.21	16.61	0.00	0.33	0.00	0.00	10.75	7.82	7.17	7.82	8.14	4.56	3.58	6.51	2.61
Gln(Q)	0.98	1.30	0.33	0.33	0.33	3.91	5.54	0.65	14.33	1.30	12.05	4.89	7.49	3.58	4.56	3.26	6.19	7.82	5.21	4.23
Arg(R)	0.00	0.65	0.65	0.33	0.00	0.98	0.65	0.98	0.98	0.00	0.65	0.33	0.98	2.28	0.65	2.93	3.58	1.95	0.98	1.95
Ser(S)	11.73	7.82	7.49	12.05	**33.22**	14.33	**24.10**	9.12	5.54	1.30	4.23	5.54	5.21	5.86	5.86	4.23	4.56	5.21	7.82	8.79
Thr(T)	4.89	4.23	5.86	4.89	5.86	8.47	6.84	4.56	3.91	0.33	3.91	10.75	11.07	13.03	6.19	9.77	6.84	7.49	7.49	6.84
Val(V)	7.82	7.49	6.84	6.19	4.23	4.56	4.56	12.38	5.86	0.00	3.26	5.86	9.77	6.51	9.45	6.84	10.10	6.84	6.19	8.47
Trp(W)	1.30	0.33	0.65	0.98	0.33	0.65	0.00	0.00	2.28	0.00	0.65	1.63	0.65	1.95	0.33	0.98	1.63	1.30	1.30	1.30
Tyr(Y)	0.33	0.33	0.00	1.30	0.33	0.00	2.93	0.00	4.89	0.00	1.95	0.33	1.95	3.58	3.58	2.28	3.58	3.58	3.91	2.61
Charged	**0.98**	**1.30**	**1.95**	**1.30**	**0.65**	**3.58**	**2.61**	**1.30**	14.01	**0.33**	17.92	36.48	16.29	22.15	18.57	28.01	22.80	20.85	19.54	25.73
Small	40.72	35.18	39.41	35.83	**68.08**	48.53	45.93	**71.99**	12.70	**97.39**	**52.12**	17.92	20.85	20.52	22.80	21.17	20.52	26.06	25.73	27.04
Aliphatic	**61.89**	**66.12**	**54.07**	49.84	40.39	31.27	26.71	**81.43**	31.60	**93.49**	48.53	16.29	30.62	27.69	34.20	25.41	30.94	28.66	25.41	29.97
Hydrophobic	45.28	48.53	37.79	**52.44**	**50.16**	39.09	44.63	**76.55**	35.18	**93.16**	49.84	29.64	42.02	34.85	41.69	40.39	44.63	37.79	37.46	43.97
Polar uncharged	27.04	20.20	32.25	27.04	46.25	48.21	49.84	16.29	34.20	**6.19**	31.92	33.88	39.74	38.76	35.83	34.53	31.92	41.37	43.97	36.16

When we measure the net charge of SPs and MPs individually (Figure 11), bacterial SPs are overwhelmingly positively charged (>0) while their MPs gravitate towards a net negative-charge bias. Median net charge for SPs_{Gram+} and SPs_{Gram-} are +3 and +2 respectively. Eukaryotes share a somewhat similar net charge distribution in their MPs when compared to $MPs_{Bacteria}$, but their SP moieties support a more uniform net charge distribution (+ve: 57.3%; neutral: 32.9%; -ve: 9.8%) in comparison to the positive-charge preference in $SPs_{Bacteria}$.

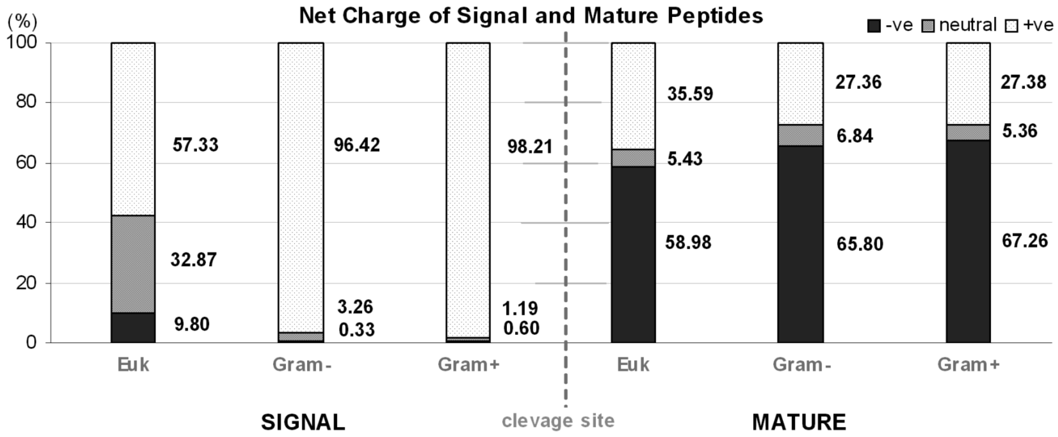


Figure 11: Net charge calculations of SPs and MPs for the three groups of organisms. The net charges are grouped into three classes: positive (>0), neutral (=0) and negative (<0) charge. The numbers represent the frequencies of which the charges are observed. The diagrams are generated using Microsoft Excel.

To examine the extent of difference in amino acid composition between the SP and MP of eukaryotes and bacteria, scatter plots are constructed for p*I*, aliphaticity, GRAVY and mean charge calculations. These features are plotted against the length of SPs (■) and MPs (▲) (Figure 12). In all three groups of organisms, the overall computed values of MPs tend to cluster in a narrower range compared to SPs. For instance, based on the calculation using aliphatic index, MPs_{Gram+} lie mostly between value of 50 to 100 whereas SPs_{Gram+} occur anywhere between 75 to 200. A similar trend exists in the other calculations including GRAVY and p*I* except for the p*I* of

MPs_{Euk}. SPs_{Euk} form two clusters based on p*I* calculation whilst SPs_{Gram+} and SPs_{Gram-} are predominantly represented within single clusters with median p*I* values of 10.3 and 10.0, respectively. From hydropathicity calculations, the GRAVY score of SPs are largely positive (SPs_{Euk}:99.7%; SPs_{Gram+}:93.5%; SPs_{Gram-}:97.7%) indicating a hydrophobic propensity while MPs show preference towards hydrophilic nature (MPs_{Euk}:93.7%; MPs_{Gram+}:94.6%; MPs_{Gram}:95.1%).

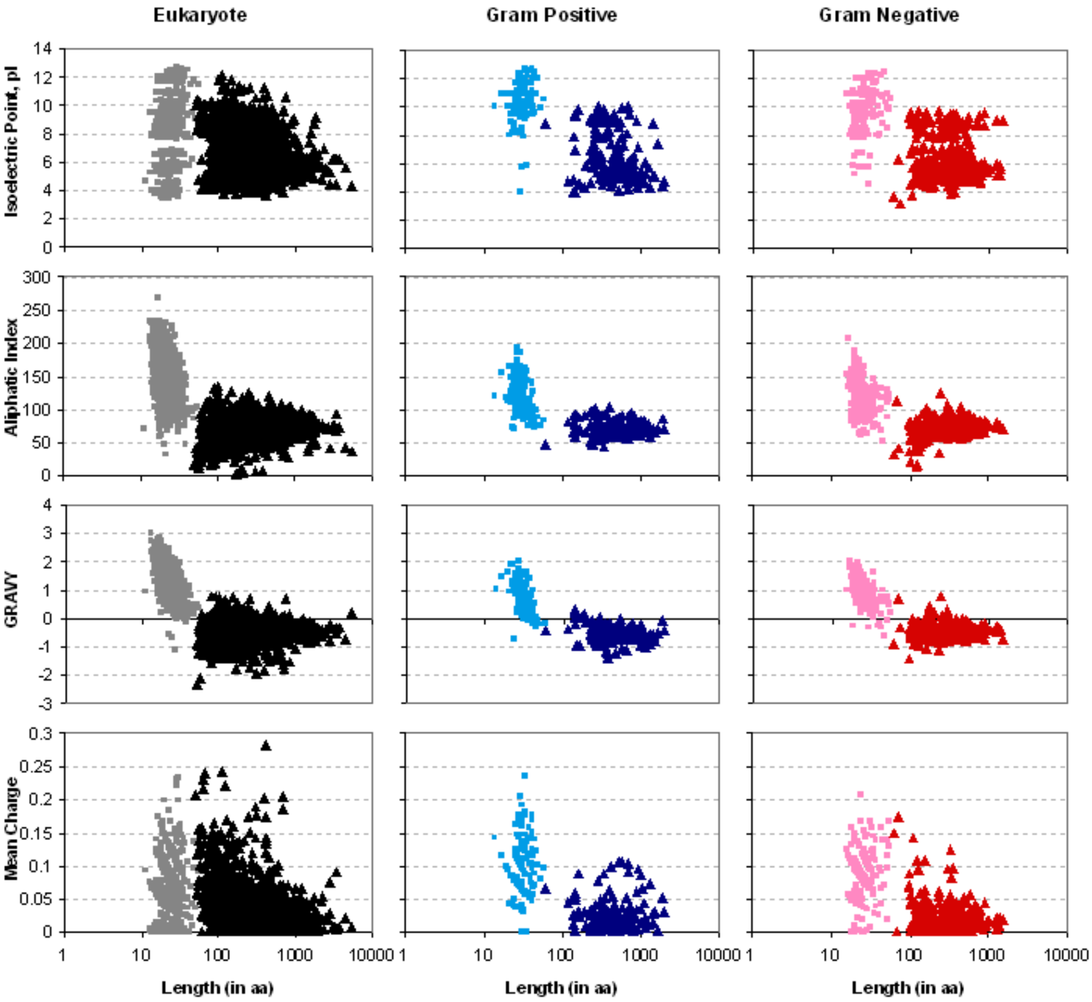


Figure 12: Comparison of the p*I*, aliphatic index, GRAVY value and mean charge among the three organism groups. Data are represented by squares (■) which denote SP while triangles (▲) denote MP.

4.4 Discussion

In this study, we have inspected a curated dataset of SP sequences to examine their variability in length and composition. We also surveyed the residues around the cleavage-processing site to locate any possible pattern. Part of the MP region is also explored since the environs of the scissile bond may provide clues to the precision in cleavage of SPs. We did not proceed further downstream of the MP region since we hypothesize that the information for cleavage processing should not be contained too far from the cleavage site, although they were studies that have proposed that the region involved could be farther downstream (Kajava *et al.*, 2000). One reason is the region is not even exposed to the cleavage machinery when it emerges from the translation process (*Chapter 2*). In addition, current prediction methods that mainly rely on SP region and possibly a few residues from the MP, are already able to predict with good accuracy of the SP (*Chapter 6*).

4.4.1 Inter-group differences

The result indicates that SPs_{Gram+} and SPs_{Gram-} share more similarities, compared to SPs_{Euk}. When the net charges of the SP of these three groups are measured (Figure 11), it is observed that SPs_{Euk} is distinctly different from the bacterial SPs in that bacterial SPs overwhelmingly favor a net positive charge bias whereas SPs_{Euk} do not exhibit such inclination. Moreover, from the constructed frequency occurrence matrices (Table 4) as well as the sequence logos (Figure 10) of these three groups, it becomes clear that the bacterial datasets indeed bear much resemblance in their overall features and properties, such as the diverse variability in their SPs primary structure, the highly-visible P3-P1 sequence motif which exhibits high selectivity for small, aliphatic residues and a detectable hydrophobic-region (*h-region*) at the core of SPs. Even so, underlying these commonalities are inter-group differences. For example, the mean length and *h-region* of $\text{SPs}_{\text{Gram+}}$ are considerably longer than those of $\text{SPs}_{\text{Gram-}}$ and SPs_{Euk}. In the case of the tripartite structure consisting of n-region (positively charged), *h-region* (hydrophobic) and *c-region* (neutral and polar) which are commonly reported in the literature, our findings show that this structure is unmistakable in the bacterial SPs but somewhat ambiguous in SPs_{Euk}, specifically in the *n-region* where positively-charged residues are far less prominent. Likewise, the sequence motif located at P3 and P1 of bacterial SPs, is largely dominated by Ala and Val, while such exclusivity is not observed in SPs_{Euk} where a number of other different amino acids are tolerated. These nuances are likely attributed to the differences in their cell-membrane structures, suggesting certain overall, minimal requirements at the sequence and possibly at structure level (Duffaud and Inouye, 1988) as well that a SP must conform to, for recognition and processing in the secretion pathway. Perhaps this may account for the selectivity for certain types of amino acids at certain subsites while simultaneously maintaining a generous accommodation for amino acid degeneracy at other subsites in the SP.

4.4.2 Influence of the mature moiety

Since the "(-3, -1) rule" (von Heijne, 1986a) was proposed, where small, uncharged residues are favored at the P3 and P1 positions, the SP has drawn much attention. A fair number of ensuing reports also began to explore the influences of the MP besides the SP, for instance, Wickner suggested that part of the translocation information is encoded within the MP while SP contains the targeting information (Wickner, 1979; Wickner, 1980). Many subsequent studies continue to furnish additional support and evidence to advance our comprehension of the less understood role of the amino acids at the MP. Numerous studies (Andrews *et al.*, 1988; Bankaitis and Bassford, 1985; Chou, 2001a; Kajava *et al.*, 2000; Le Loir *et al.*, 2001) experimented with SPs by fusing them to an assortment of secretory and non-secretory proteins for homologous and heterologous secretion and demonstrated that the SP alone is not sufficient to ensure the processing of secretory proteins, implying that a section of the MP must contribute to the process. In fact, such studies have shown that a balance between the SP and portion of the MP affects export efficiency (Li *et al.*, 1988; Summers *et al.*, 1989; Summers and Knowles, 1989). In an analysis of the interactions of SPs from different substrates with the translocation channel components, it was suggested that certain arrangement or pairing of the SP with their natural MP may confer translocation-competent conformation which may not be properly achievable in a heterologous context, thus arguing for the influence of the MP (Kim *et al.*, 2002).

Examination of the frequencies between the adjacent positions of 10aa from both sides of the cleavage site (Table 4) viz. SP (P10 to P1) and MP (P1'to P10') for all three organism groups reveals that the frequencies of charged residues (counting both positively and negatively charged residues) are relatively stable. The transition value from one position to another does not fluctuate beyond 50% of the difference for the MP. For the SP moiety (P10-P1), the fluctuations are more dramatic at P5, P4 and P2 (although less pronounced for Gram- bacteria) while virtually absent at other positions. When the charged residues are divided into positively and negatively charged subgroups, it is observable that a specific charged subgroup is preferred at certain positions. Moreover, when the mean charge is measured using a sliding window of variable size (3 to 11), the fluctuations between the positively and negatively charged residues seem to converge and stabilize at around P8’ to P10’ whereas uncharged residues maintain a uniform trend throughout all the positions.

Approximately a quarter of the bacterial MPs and 35% of MPs_{Euk} bear a net positive charge, 5-6% are neutral while the majority of MPs favor a net negative charge. This is in stark contrast to the SP moiety that is inclined towards a net positive charge, the trend being especially strong in bacteria. Probably, secretory proteins maintain their desired net charge levels within the SP and MP to enable their interaction with other players in the secretion pathway. This can be done by varying or accommodating diverse amino acids at selected positions while being rigid in the choice of amino acids at others. This selectivity is visible at some MP positions particularly those in the vicinity of the cleavage site but not further downstream.

It was proposed that a net charge with null or negative bias should be maintained for the first eighteen amino acid residues of the MP to promote successful expression of proteins in Gram- bacteria and any optimization performed on the SP should include the specified region (Kajava *et al.*, 2000). In this study, no significant pattern is observable beyond P5’ at the MP based on the results (Table 4 and Figure 10) to support this proposal, possibly because the first eighteen residues could include several combinations of SPs and MPs. Moreover, the relative frequencies of adjacent positions at the MP appear to be rather stable. The results here are generally in agreement with other studies that included the MP, but the extent of the region to be included remains debatable. The varying results from the different studies make it difficult to compare and obtain consensus. Furthermore, the paucity of crystal structures solved to date (only four SPase I-related entries are found in the PDB (Berman *et al.*, 2000)) adds to the challenge of deciphering the extent of MP involvement in the secretory pathway.

4.4.3 Recognition of the cleavage site and its flanking region

Based on the dataset used in this study (1,877 eukaryotic, 168 Gram+ and 307 Gram- sequences), the frequency of occurrence of the canonical sequence motif Ala-X-Ala (von Heijne, 1986a), which is in fact (small and aliphatic residue)-X-(small and aliphatic residue) (see Section 2.3.4 for the detailed description of residues that are represented by this motif), at P3 and P1, appear with frequencies of 61.6%, 61.9% and 77.5% respectively. From these frequencies and Table 4, it is clear that there are sequence patterns (if any) that do not conform to this pattern, implying that the sampling space for cleavage site recognition is not limited to the canonical Ala-X-Ala motif. This is further supported by the observations of residues which were forbidden to occur at these positions according to the "(-3, -1) rule" (von Heijne, 1986a).

We also observed several prominent patterns upon scrutinization of these flanking residues. Pro has been implicated as a structure disruptor due to its steric hindrance from its cyclic side chain and inability to form a hydrogen bond that stabilizes a helix (Martoglio and Dobberstein, 1998). Pro is often found at the end of α-helices, in turns or loops but produces a bend when it appears in the middle of an α-helix. Pro is markedly disfavored from P3 to P1' but it is comparatively prevalent at P4 and P2' (Table 4). The turn-inducing Pro is supported by studies that have shown the presence of a β-turn at the P5 to P1 region of SPase-substrate complex (Karamyshev *et al.*, 1998). On the other hand, the absence of Pro (particular in bacterial SPs) at P3 to P1' is consistent with reports on impaired function or inhibition of SPase I with Pro appearing at these positions (Barkocy-Gallagher and Bassford, 1992; Nilsson and von Heijne, 1992). Glycine, another helix-breaking residue, is also spotted in modest amount at P5 and P4. As we have seen earlier, the canonical Ala-X-Ala sequence motif for the SP cleavage site accounts for only approximately half of the recognition sites. By considering these flanking residues, many non-canonical cleavage sites can be considered. These features may possibly work in concert to provide the secretory machinery flexibility, versatility and perhaps accuracy to enact the SP recognition processes.

4.5 Summary

A total of 2,352 SP-containing sequences are assembled from a variety of organisms using an improved protocol (*Appendix B*). Sequences are analyzed on their physico-chemical properties such as p*I*, aliphatic index, GRAVY score, hydrophobicity, net charge and position-specific residue preferences. Findings from the analyses of these sequences show that the eukaryote, Gram+ and Gram- groups share several similarities in general but they display distinctive features as well, in terms of their amino acid compositions and frequencies, and physico-chemical properties. Additionally, the physico-chemical properties can be used to identify spurious sequence entries that purportedly contain SP, thus adding another method for error detection in our semi-automated pipeline.

When we inspect the sequences, we observe certain incidences of residues such as turn-promoting residues at the flanking regions of the cleavage site. Furthermore, there are also slight patterns of residues that occur downstream of the cleavage site. These flanking residues may influence the cleavage processing and contribute to non-canonical cleavage sites.

In spite of these patterns including the canonical motifs, these observations are unable to account for other SPs which do not bear such resemblance at their cleavage site. Furthermore, studies have shown that introducing or replacing the original residues at positions such as P3 to P1 may result in alternative cleavage sites (Fikes *et* *al.*, 1990). In this regard, how does the machinery recognize these non-canonical sequences? Moreover, the preference for certain residues or perhaps certain 'types' of residues at specific positions suggests that there may be other features that are involved which are not captured or manifested by the sequences. This notion has been suggested before whereby there may be certain minimal requirements at the sequence and possibly at structure level that a SP must conform to, for recognition and processing in the secretion pathway (Duffaud and Inouye, 1988). It was even suggested that a 'precise or right' alignment may be needed of the SP for the cleavage event to occur (Jain *et al.*, 1994). Thus, it will be interesting to investigate these regions from the structural perspective.

Chapter 5: Structural Analysis of SPs

5.1 Introduction

Chapter 2 describes a host of proteins/RNAs molecules that interact with SPs in the secretion pathway. One interaction partner is the SPases I that play essential roles in the viability of bacteria (Date, 1983; Klug *et al.*, 1997). These enzymes are responsible for the cleavage of SP from proteins that are translocated across biological membranes (*Chapter 2*). Until now, the crystal structure of SPase I in complex with SP has not been solved. The worldwide archive of biological macromolecules — PDB, contains only four structures related to SPases I, and they all belong to *E. coli*.

Currently, *E. coli* is by far the most widely studied and used host organism for the bacterial expression of heterologous secreted proteins, especially for therapeutic purposes, with reported yields of 5–10 g/L (Georgiou and Segatori, 2005). Mutations in SP have been known to affect secretion either by enhancing the processing of the cleavage site or by inhibiting this proteolytic processing (Martoglio and Dobberstein, 1998). It is known that besides the SP, the N-terminus region of the MP is known to affect protein secretion (Andersson and von Heijne, 1991).

The *E. coli* SPase I is of particular interest in the study of SPases I, as its active site is relatively accessible at the bacterial membrane surface (Paetzel *et al.*, 2000; Wolfe *et al.*, 1983). Although many mutational and biochemical studies have been performed, basic questions such as SPase I fidelity and substrate specificity remain unanswered. SPs exhibit limited primary sequence homology, but they are well conserved at P3 to P1 relative to the cleavage site (Fikes *et al.*, 1990). Comparative analysis of thirty-six prokaryotic signal peptides reveals that SPases I specifically recognizes substrates with small neutral residues at both the P3 and P1 positions (von Heijne, 1986a). P3 is dominated by the presence of Ala, Gly, Ser, Thr and Val; while P1 is characterized by Ala, Gly, Ser and Thr (Fikes *et al.*, 1990; von Heijne, 1986a). Accordingly, the P3 and P1 positions have been proposed to constitute the SPase I cleavage site and have been actively applied by various groups for predicting SP cleavage sites (Fikes *et al.*, 1990; Folz *et al.*, 1988). These findings are cited as affirmation of the location of two key determinants within the SP cleavage site. Unfortunately, no solution structures exist that can illustrate precisely how the precursor protein is oriented within the SPase I substrate binding site prior to proteolysis, or the identity of other critical determinants that control substrate specificity (Karla *et al.*, 2005). In *Chapter 4*, the SP sequences have been analyzed on a large-scale basis and some of these curated sequences can be used in this work. The aim is to understand the determinants that are involved in SP recognition, binding and cleavage, from a structural viewpoint.

This chapter reports our findings from the modeling of an *E. coli* periplasmic disulfide-bond A oxidoreductase (*DsbA*) 13-25 in complex with its endogenous SPase I based on the crystal structures of *E. coli* SPase I in complex with β-lactam (Paetzel *et al.*, 1998) and lipopeptide (Paetzel *et al.*, 2004) inhibitors. The *DsbA* 13-25 precursor protein was selected for this study for its efficient periplasmic secretion (Karla *et al.*, 2005). By threading the P7 to P1' positions against the solved structures of β-lactam (Paetzel *et al.*, 1998) and lipopeptide (Paetzel *et al.*, 2004) inhibitors, this newly generated model reveals that precursor protein is bound to *E. coli* SPase I with a pronounced twist between positions P3 and P1'. Thirteen subsites S7 to S6' that might be critical to these and other aspects of catalysis are identified. This model is further corroborated by comparative analysis of one hundred and seven experimentally validated substrates taken from the set described in *Chapter 4*.

5.2 Materials and Methods

5.2.1 Preprotein sequence data

107 preprotein sequences are extracted from the SPdb database to be used as the substrates for *E. coli* SPase I (Choo *et al.*, 2005). Redundancy reduction is performed on these sequences where sequences with 80% sequence identity are removed using the CD-HIT software (Li and Godzik, 2006).

5.2.2 Crystallographic data

The atomic coordinates of *E. coli* SPase I are extracted from the PDB entry 1B12 (Paetzel *et al.*, 1998) which has a 1.95 Å resolution structure. Atomic coordinates for *E. coli* SPase I-bound *β-lactam* (Paetzel *et al.*, 1998) and lipopeptide (Paetzel *et al.*, 2004) inhibitors are retrieved from PDB entries 1B12 and 1T7D respectively. The structures are relaxed by means of conjugate gradient minimization, using the Internal Coordinate Mechanics (ICM) software package (Abagyan *et al.*, 2004).

5.2.3 Substrate modeling

Protein threading or fold recognition (Akutsu and Sim, 1999) computes an alignment between a target sequence and the template structure using a scoring function where the best-fit spatial positions of the template structure are used to construct the structural model for the target sequence. The coordinates for P7 to P1' of *DsbA* 13-25 are obtained by threading against the crystallographic structures of *E. coli* SPase I-bound inhibitors (Paetzel *et al.*, 1998; Paetzel *et al.*, 2004). Coordinates for P7 to P3 were taken from the structure of *E. coli* SPase I-bound lipopeptide inhibitor (Paetzel *et al.*, 2004) by substituting the locations of atoms N1, C2, C5, O6, N7, C8, C10, O11, N12, C13, C14, O15, N16, C18, C26, O27, N28, C29, C31, O32, and N33 with DsbA 13-17 main-chain atoms; while coordinates for P2 to P1' are guided by the solution structure of the *E. coli* type I SPase-bound β-lactam inhibitor based on the location of atoms N4, C5, C6, C3, C9, O10, C15, C16, O17, C18, O19, C20. A flexible docking using biased Monte-Carlo procedure (Abagyan and Totrov, 1999; Abagyan *et al.*, 2004; Paetzel *et al.*, 1998) (Fernandez-Recio *et al.*, 2002) that incorporates the "Rapid Exact-Boundary Electrostatics" algorithm for evaluation of the electrostatic solvation energy (Totrov and Abagyan, 2001) is subsequently performed to sample the different positions and orientations of P2' to P6' with respect to the receptor. In each iteration, a random move in the P2' to P6' of the ligand is performed and new conformations are selected based on the Metropolis criterion with a temperature of 5000K (Fernandez-Recio *et al.*, 2002; Metropolis *et al.*, 1953). The simulation is terminated after twenty thousand energy evaluations (Fernandez-Recio *et al.*, 2002) and the results are analyzed for consistency.

5.2.4 Intermolecular hydrogen bonds

The number of intermolecular hydrogen bonds is calculated using HBPLUS (McDonald and Thornton, 1994) in which hydrogen bonds are computed according to the criteria if (i) it is between a listed donor and acceptor (ii) the angles and distances formed by the atoms surrounding the hydrogen bond lie within the set criteria. Further details of the calculation can be found within the user manual of HBPLUS (http://www.csb.yale.edu/userguides/datamanip/hbplus/).

5.3 Results and Discussion

5.3.1 Substrate binding site

The energetically favored and most frequently populated bound conformation of DsbA 13-25 H2N-LAFSASAΔAQYEDG-COOH, where the cleavage site is indicated by Δ (Perna *et al.*, 2001), to *E. coli* SPase I is obtained from the generated structural model. The complex defines thirteen enzyme subsites, S7 to S6', within the SPase I substrate binding site, interacting with bound precursor SPases I. Among these, six smaller clefts or 'pockets' are identified at subsites S3, S2, S1, S1', S3' and S4' respectively (Figure 13). The narrow clefts at S3, S2, S1 and S1' play direct roles in the high specificity of the SP residues while the larger clefts at S3' and S4' may be responsible for the specificity of the mature moieties.

The side chain of Ala19 (P1; Figure 13a) is buried within the S1 subsite, which is composed of primarily hydrophobic and non-polar enzyme residues including the previously identified Ile86, Pro87, Ser88, Ser90, Met91, Leu95, Tyr143, Ile144, and Lys145 (Paetzel *et al.*, 2000; Paetzel *et al.*, 2002a). The S2 subsite (Gln85, Ile86, Pro87, Ser88, Met91, and Ile144) constitutes the deepest cavity within the substrate-binding site. This pocket can accommodate residues with large side chains and appears to play an important role in substrate specificity of *E. coli* SPase I, consistent with biochemical experiments (discussed in *Substrate specificity*). This subsite, formerly proposed as the S1 subsite by (Paetzel *et al.*, 2000; Paetzel *et al.*, 2002a), largely overlaps with the S1 subsite due to a pronounced twist in the P3 to P1' binding conformation (Figure 14; detailed in *Substrate binding conformation*). This model reveals that Ser18 (P2) side chain is not solvent exposed but it is completely buried at this location due to a pronounced twist in the P3 to P1' binding conformation

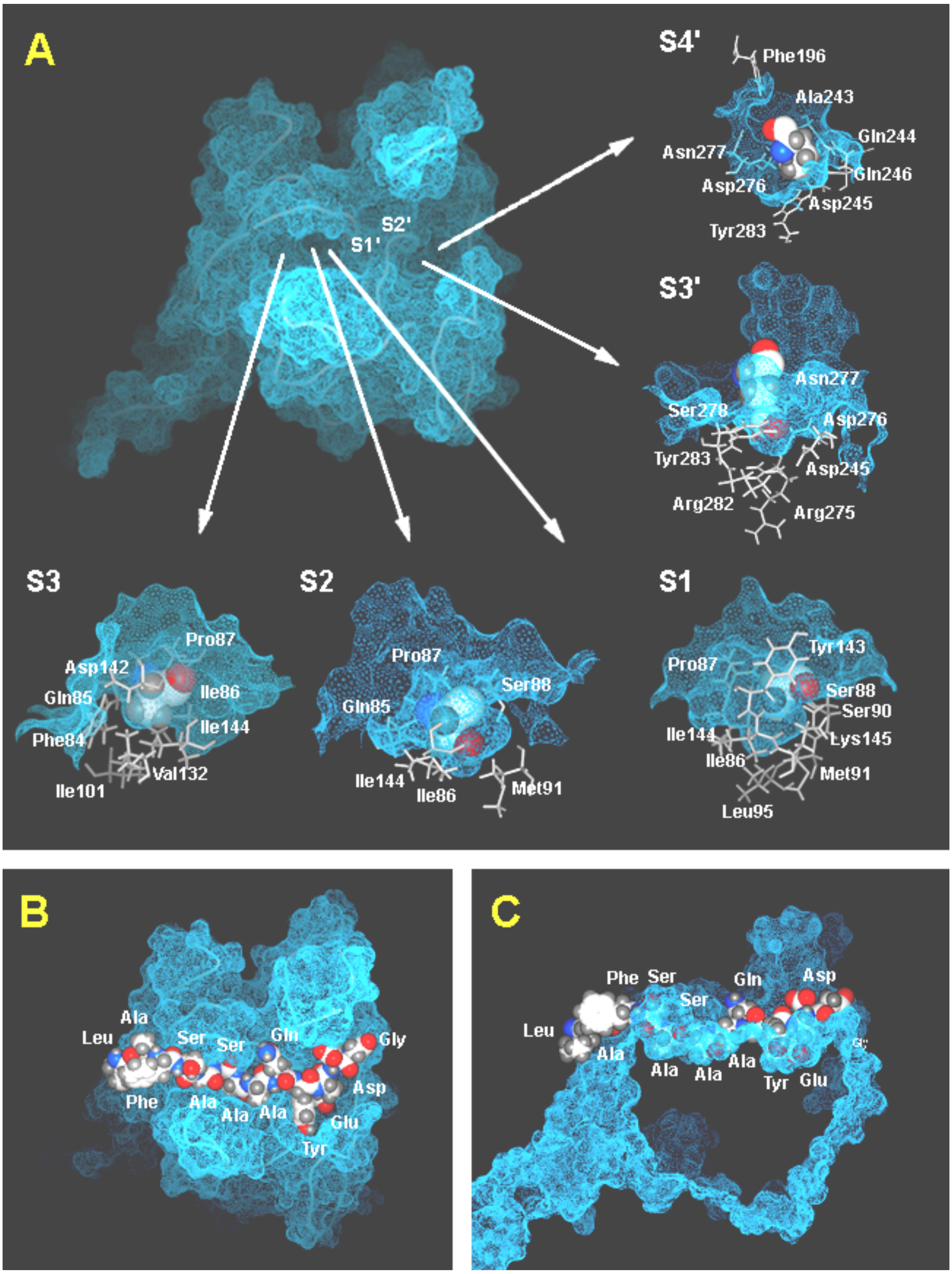


Figure 13: The *E. coli* SPase I substrate binding site. Pockets defining the binding site of *E. coli* SPase I. A) Top view of the molecular surface of E. coli SPase binding site (colored blue) with Cα trace of SPase (blue lines). Pockets that accommodate SP side chains are shown in detail in surrounding views and numbered in accordance to their position along the peptide from the S1 pocket that contains the active-site nucleophile, Ser90. B) Top view of the molecular surface of E. coli SPase binding site (colored blue) with the bound conformation of DsbA precursor peptide as a CPK model. C) Side view of structure in B, rotated by 90°. The structures are generated using the ICM modeling software by Abagyan *et al.*, 2004.

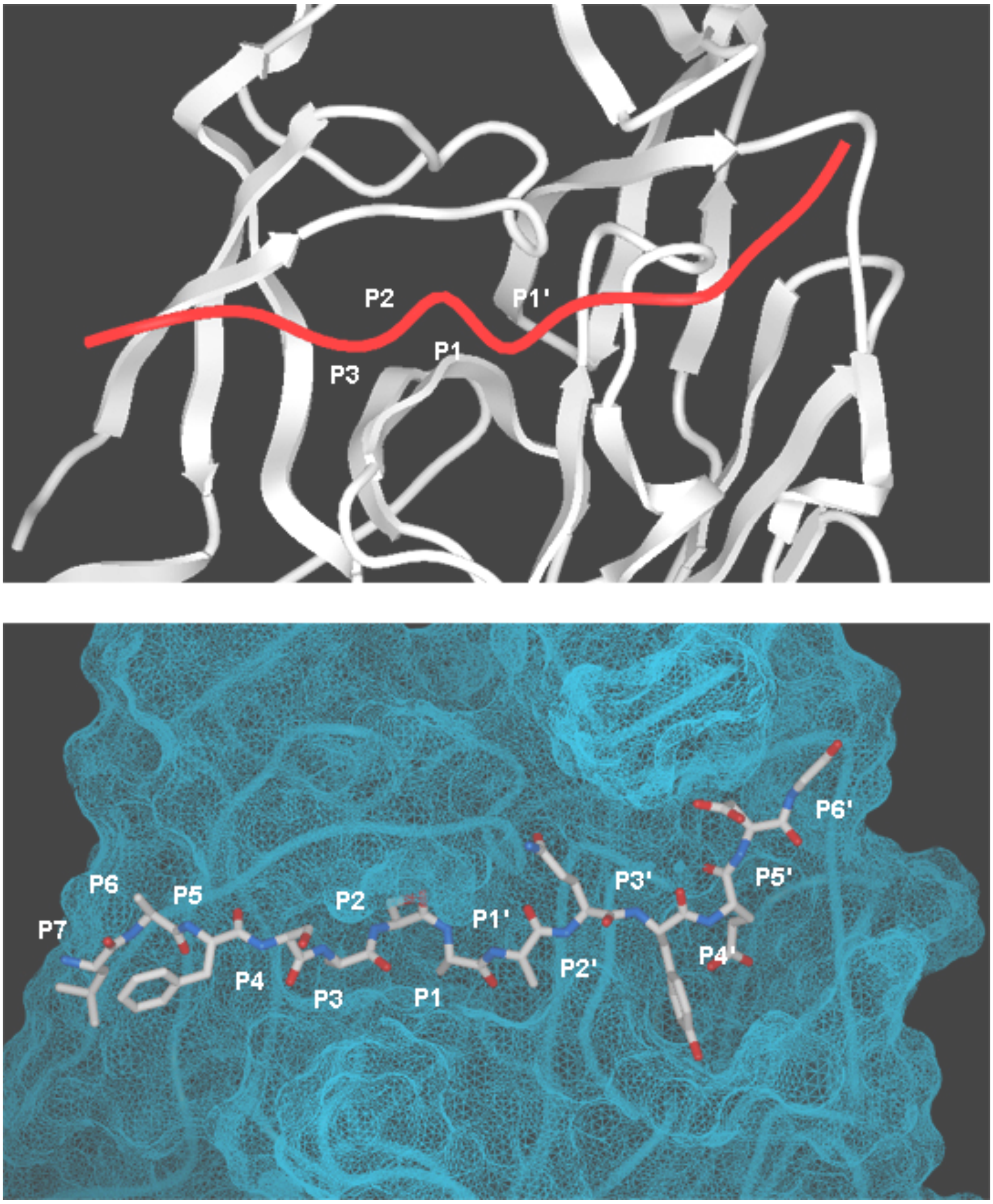


Figure 14 A model of the DsbA 13-25 precursor protein (Cα trace in black) bound to the active site of *E. coli* SPase I (schematic ribbon diagram in gray) illustrating a pronounced twist in the peptide backbone between P3 and P1’ at the catalytic site.

(Figure 14; detailed in *Substrate Binding Conformation*). The S3 subsite, which is composed of non-polar atoms from residues Phe84, Gln85, Ile86, Pro87, Ile101, Val132, Asp142, and Ile144 (Paetzel *et al.*, 2000; Paetzel *et al.*, 2002a), is located diagonally across from the S1 subsite. This pocket constitutes the third deepest cavity within the substrate-binding site and can accommodate a wide variety of side chains. The S4 subsite, consisting of Phe84, Gln85, Pro87, and Asp142, is in contact with Ser16 (P4). Further upstream, the S5 subsite is defined by Phe84, Gln85, and Asp142; S6 consists of Pro83 and Phe84; while the S7 subsite includes Glu82 and Pro83.

At P1' to P6' of the mature moiety, this model indicates that the side chains of P1' to P5' residues are in position to make significant contact with the *E. coli* SPase I. The S1' subsite shares similar residues with the S1 subsite and includes Ser88, Ser90, Tyr143, and Ala279. The S2' subsite includes Ser88, Ser90, Phe208, Asn277, and Ala279. The S3' and S4' subsites constitute a broad pocket that can accommodate both positive and negative charged residues by re-arrangement of side chains (Figure 15). The S3' subsite is composed of Met249, Tyr50, Asp276, Asn277, Ala279, Arg282, and Tyr283, while the S4' subsite includes Gln244, Asp245, Asp276, Asn277, and Arg282. Further downstream, the S5' subsite consists of Phe196, Ser206, Ala243, Asp276, and Asn277, while the S6' subsite includes Phe196, Ile242, and Ala243.

In this model, the bound precursor protein makes significant contact with *E. coli* SPase I from S7 to S6'. Models described earlier focused solely on the P3-P1' region and did not analyze in full the different substrate binding pockets on either side of the scissile bond. In particular, the S2 subsite is formerly proposed as the S1 subsite (Paetzel *et al.*, 2000; Paetzel *et al.*, 2002a), as it largely overlaps with the latter. In contrast to the analysis by (Paetzel *et al.*, 2002a), this model reveals that the Ser18 (P2) side chain is not solvent exposed but is completely buried at this location. The ability of S3'/S4' to alter their electrostatic requirements by varying side chain conformations (Figure 15) may help explain the propensity to find substrates with charged amino acids at these positions (discussed in detail in *Substrate Specificity*).

5.3.2 Substrate binding conformation

This newly generated model is constructed by using the coordinates of P7 to P1' and threading the region against the solved structures of β-lactam (Paetzel *et al.*, 1998) and lipopeptide (Paetzel *et al.*, 2004) inhibitors in complex with *E. coli* SPase. This is followed by *ab initio* docking of P2' to P6' (details described in *Methods*). The precursor protein, DsbA 13-25 is bound to *E. coli* SPase I in an extended conformation with a pronounced backbone twist between Ala17 (P3) and Ala20 (P1') (Figure 14). In the P3-P1' segment, the first three side chains are oriented towards the binding groove while the P1' side chain is oriented across the binding groove. As shown in Figure 16, similar interactions between the *E. coli* SPase with DsbA 13-25 model, lipopeptide inhibitor (PDB ID: 1T7D) and β-lactam inhibitor (PDB ID: 1B12) are observed. The conformations of P3' and P4' allow their corresponding side chains to extend into a large cavity (S3'/S4' subsite; Figure 13). As such, medium or large residues are preferred at these two positions for favorable interactions. Good agreement with the known experimental data (refer to *Substrate Specificity*) is obtained, supporting the validity of our model.

Ten positions for hydrogen bonding were identified supporting high affinity binding between *E. coli* SPase I and DsbA 13-25. These include Ser18 (P2) O…Ser88 NH, Ser18 (P2) O…Ser88 OG, Ala19 (P1) N…Ser88 OG, Gly89 N…Gln21 (P2') OE1, Ala19 (P1) N…Ser90 OG, Ser90 OG…Ala20 (P1') O, Lys145 Nζ…Ala19 (P1) O, Gln194 NE2…Asp24 (P5') OD2, Ser206 OG…Asp24 (P5') OD2, and Arg282 NH1…Glu23 (P4') OE1. Our model suggests that the enzyme-substrate contact points extend all the way from P7 to P6' of the *DsbA* precursor protein.

The orientation of *DsbA* 13-25 side chains within the active site (P7-P6’) of *E. coli* SPase I adopts the pattern (Tong *et al.*, 2004): ↓•••↓↓↓••↓↓•• (where ↓ represents a side chain oriented towards the binding site and • represents a side chain oriented away or across the binding site). Specifically, the P3-P1’ portion adopts the pattern: ↓↓↓•, with the side chains of P3, P2 and P1 oriented towards the binding groove thereby supporting the stringent selectivity criteria in this region. The side chain of P1’ alone is oriented differently, in accord with the observed variability in this position. A similar conformation was obtained for the precursor sequence OmpA 15-27 (Carlos *et al.*, 2000; Ekici *et al.*, 2007) H_2N-FATVAQAΔATSTKK-COOH (P1-P1’ cleavage site indicated by Δ) in complex with *E. coli* SPase I (data not shown). Here again, the P3-P1’ side chains of OmpA adopt the orientation ↓↓↓•, while the model proposed by (Paetzel *et al.*, 2002a) and (Ekici *et al.*, 2007), adopts the pattern ↓•↓•, with the side chain of P2 not pointing towards the binding groove. The disparity between this model and the model by (Paetzel *et al.*, 2002a) may be attributable to the selection of different template structures where the structures of the covalently bound peptide inhibitor complex and the analogous enzyme LexA were used to guide the P1 and P3 to P6 positions of the later (Paetzel *et al.*, 2002a), while the coordinates of P7 to P1’ for this model are guided by the solved structures of β-lactam (Paetzel *et al.*, 1998) and lipopeptide (Paetzel *et al.*, 2004) inhibitors in complex with *E. coli* SPase. For this model, the P2 side chains in the bound *DsbA* and OmpA models are hydrogen-bonded to the catalytically important SPase I residue, Ser88 (Paetzel *et al.*, 2004). The twist in the backbone conformation in the region P3-P1’ is representative of the transition state, with three critical hydrogen bonds conserved between this model and the bound β-lactamase and lipoprotein inhibitors, with the atoms Ser88 Oγ, Ser90 Oγ and Lys145 Nζ important for catalytic activity.

5.3.3 Substrate specificity

One interest of this study is to understand how the peptides modeled in this study reflect the *E. coli* repertoire of secreted signals. Comparative analysis of one hundred and seven experimentally determined *E. coli* SPase I substrates (Figure 17) revealed high conservation of amino acid residues at positions P3 and P1. In particular, P1 is dominated by small (99%), hydrophobic (98%), and neutral (100%) residues. Ala is the predominant residue (92% or 98/107) at this position, followed by Gly (9%). Position P2 shows a strong preference for bigger side chains with 87% (93/107) possessing medium- or large-size residues at this location. Position P3 also shows a preference for hydrophobic residues (83%). Although this position contains mainly small amino acid residues (61%), it can also accommodate both medium (25%) and large (14%) residues. Only 50% (54/107) of the sequences contain the consensus Ala-X-Ala recognition motif, while even fewer sequences (18%; 19/107) contain a Val-X-Ala recognition sequence. In this newly modeled structure for *DsbA* propeptide, the side chains from P7 to P4 are also in positions to make substantial contacts with SPase I (Figure 14), but are not confined to 'pockets'. Nonetheless, these residues may also participate in binding by interacting with surface residues of SPase I. These observations are in accord with the lack of residue preference observed in these positions (Figure 17). Overall, in our dataset, neutral residues (≥ 98%) are preferred in positions P7 to P1, indicating that charged interactions between SPs and *E. coli* SPase I are disfavored in this stretch, consistent with earlier reports on the carboxy-terminus of the C-region (Paetzel *et al.*, 2001). However, few SPs possess polar residues at the C-region (P7: 17%; P6: 48%; P5: 30%; P4: 52%; P3: 18%; P2: 41%; P1: 2%), in contrast with earlier studies (Paetzel and Strynadka, 2001; van Roosmalen *et al.*, 2004). Most residues are well tolerated at P1', except for Pro, Arg and large hydrophobes (Ile, Met, Trp). Here, Pro is disfavored as the rigid positioning of its backbone hinders docking interactions with SPase at P2’ to P6’. Majority of *E. coli* SPs contain medium or large residues at both P3’ (81%) and P4’ (90%). The propensity for negatively charged residues to occur at P3’ and P4’ are low, with observed values of 10% (or 11/107) and 19% (or 20/107) respectively, while 8% (9/107) and 13% (14/107) respectively of the sequences analyzed have positively charged residues at these positions.

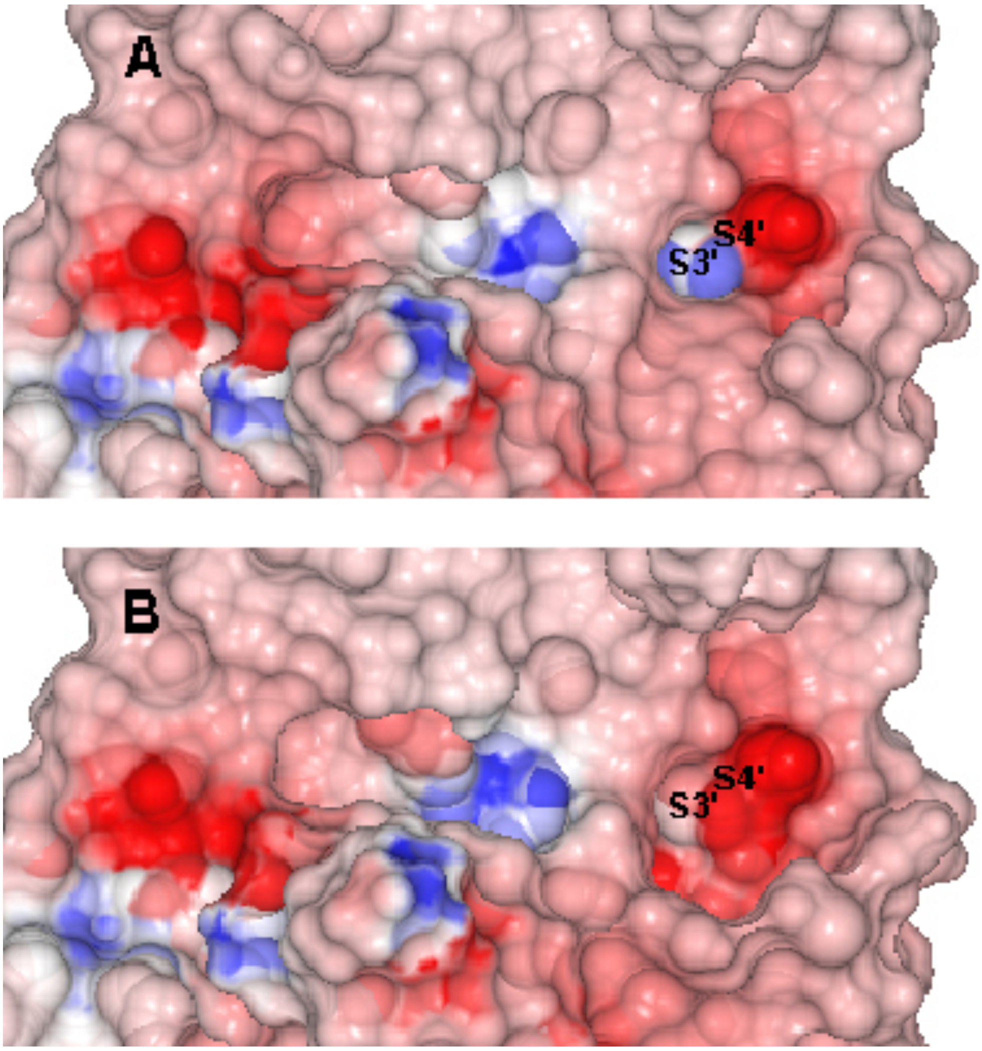


Figure 15: The S3’/S4’ subsites of *E. coli* SPase I. Rearrangements of side chain residues at S3’/S4’ subsites in the crystallographic structure of *E. coli* SPase I (PDB ID: 1B12). (A) The side chain of Asp276 is exposed to interact with amino acid residues at P3 and P4. (B) Rearrangements of Asp276 and Arg282 result in a positively charged pocket at S3’/S4’ subsites.

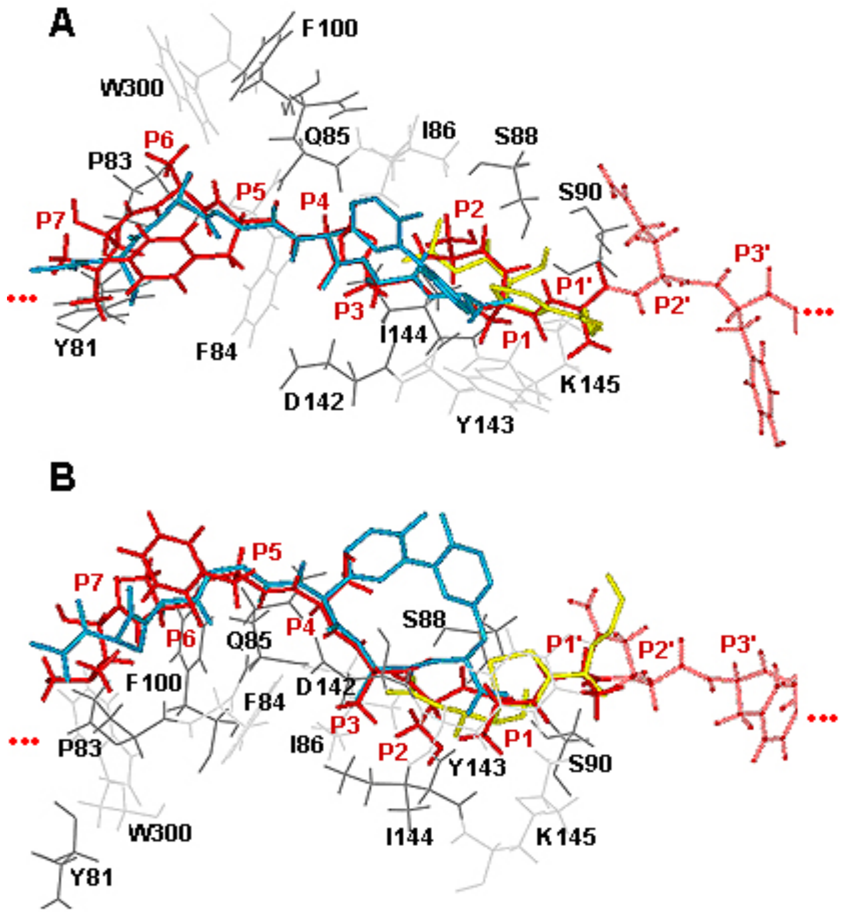


Figure 16: Superimposition of DsbA 13-25 precursor protein with lipopeptide and β-lactam inhibitors. A model of the DsbA 13-25 precursor protein (red) bound to the active site of E. coli SPase I (gray). Superimposition of the P7 to P1’ of DsbA precursor protein with the lipopeptide (blue; PDB ID: 1T7D) and β-lactam (yellow; PDB ID: 1B12) inhibitors from (A) top view and (B) side view respectively. Residues N-terminal to P7 and C-terminal to P2’ have been truncated for clarity.

Bacterial SPase I uses a Serine/Lys catalytic dyad mechanism (Paetzel *et al.*, 2000). Ser-90 acts as the nucleophile while the proposed Lys-145 constitutes the general base, working together to form an acyl-enzyme complex intermediate. Three conserved waters were observed in the SPase I apo-enzyme crystal structure (Paetzel *et al.*, 2002a) in which the 2nd water is coordinated to the Ile144 NH backbone while the 3rd water is coordinated to Lys145 Nζ. The 3rd water is proposed as the deacylating water in the SPase I catalysis. In the β-lactam acyl-enzyme structure (Paetzel, *et al.*, 1998), the 2nd and the 3rd waters are displaced by the β-lactam inhibitor whereas a recent resolved structure reported the displacement of the 2nd water (Luo, *et al.*, 2009). Thus, it is highly plausible that these displacements might play a critical role in the peptide-enzyme interaction.

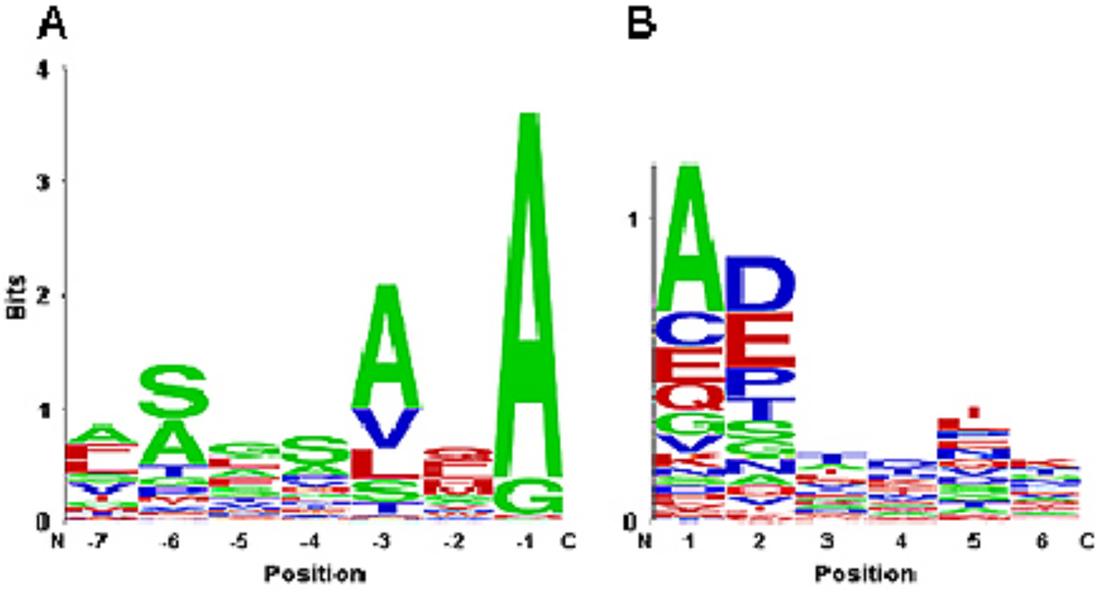


Figure 17: Analysis of *E. coli* SPs. Sequence logo illustrating the size (small: green; medium: blue; large: red) of amino acids at different positions along the precursor proteins of 107 experimentally verified *E. coli* SPs from SPdb, showing (A) the end of the SP (P7 to P1) and (B) the start of the mature moiety (P1' to P6'). Cleavage site is situated between -1 and +1.

5.4 Summary

A theoretical structural model of was created in this study by means of threading and homology modeling to model the *E. coli* periplasmic disulfide-bond A oxidoreductase (DsbA) 13-25 in complex with its endogenous SPase I based on the crystal structures of *E. coli* SPase I in complex with β-lactam (Paetzel *et al.*, 1998) and lipopeptide (Paetzel *et al.*, 2004) inhibitors for P7 to P1'. This was followed by *ab initio* docking to generate the conformations for P2' to P6'. The resulting 3D model provides an opportunity to examine the bound structure of *E. coli* SPase I complex that have been difficult to solve experimentally.

From the model, the existing and newly identified substrate binding sites provide clues to the SPase I cleavage fidelity and substrate specificity. These sites are consistent with existing biochemical results and solution structures of inhibitors in complex with *E. coli* SPase I (Paetzel *et al.*, 1998; Paetzel *et al.*, 2004). The structural analysis results correlates well with the sequence analysis presented earlier (*Chapter 4*). Several positions exhibit preference and aversion for certain types of residues at various positions. For instance, small-size residues are preferred at P3' and P1'. This is consistent with the requirement imposed by the binding groove for the SP to fit in. There is also the existence of an extended conformation of the precursor protein with a pronounced backbone twist between P3 and P1' adjacent to the cleavage site. The newly defined subsites, S1' to S6' play critical roles in the substrate specificities of *E. coli* SPase I (Karla *et al.*, 2005).

This work advances our understanding of the molecular mechanism governing SP specificities and SPase I fidelity, and can be useful in guiding the design of suitable SPs and MPs for enhancing heterologous protein expression using *E. coli* as the host organism. This knowledge will be immensely useful in aiding the development of prediction method for the SP cleavage site, which shall be explored in the next chapter. Investigation of the sequence and structures in this work support the suggestion of other experimental studies that the SP and MP play direct role in catalysis, thus they should be considered during the development of predictive tools.

Chapter 6: Computational Prediction of SPs

6.1 Introduction

Several lines of work have investigated different aspects of targeting signals, including the determination of the targeted cellular localization upon translocation and the identification of efficient signal sequences. One particular work that interests us is the identification of SP and its cleavage site. This work is fundamentally important as it impacts on other features such as transmembrane topology (Reynolds *et al.*, 2008), subcellular localization (Emanuelsson *et al.*, 2000; Bodén and Hawkins, 2005), structure modeling and prediction (Kanagasabai *et al.*, 2007), assignment of putative functions to novel proteins and identification of putative cleavage sites in database annotation (Menne *et al.*, 2000), to name a few examples. Importantly, the systematic functional annotation of biological sequences using Gene Ontology (GO) (Ashburner *et al.*, 2000) requires a precise knowledge of the subcellular localization, where SP prediction has a fundamental input.

Moreover, the vast numbers of unprocessed sequences that are deposited continually into the public databases require rapid functional annotation techniques, with subcellular localization being a key feature. Rising industrial demand further presses for more effective methods to raise expression levels in recombinant systems. Consequently, these factors have catalyzed the development of a myriad of computational methods to automate SP prediction (Table 5), ranging from simple weight matrices to sophisticated machine learning algorithms. Machine learning techniques are particularly popular, and they are especially useful in domain where sequence data abound but our working knowledge of the underlying mechanism is limited. Their robustness to ‘noise’ in data enables them to achieve better accuracy

Table 5: Software tools that are publicly available for the prediction of SPs (includes the detection of SP and its cleavage site). Tools/methods which have been discontinued from development or unavailable for use are omitted. A comprehensive and updated listing of databases and prediction tools related to protein targeting or sorting is available at (http://www.psort.org/). Abbreviations used in this table (HMM= Hidden Markov model; ANN= Artificial neural networks; OET-KNN: Optimized evidence-theoretic K-nearest neighbor; PWMs=Position weight matrices; SVM=Support vector machines).

Name	Method type	Dataset division	Description (website URL)
Philius (Reynolds *et al.*, 2008)	Dynamic Bayesian Networks	No division	Inspired by Phobius, this tool is also designed for transmembrane protein topology prediction. It is capable of predicting SPs as well since it incorporates a SP submodel in addition to a transmembrane submodel. Training data from Phobius (Käll *et al.*, 2004) is used. (http://www.yeastrc.org/philius/pages/philius/runPhilius.jsp)
Phobius (Käll *et al.*, 2004)	HMM	No division	A combined predictor for transmembrane protein topology and SP where the different regions of transmembrane and SP are modeled respectively. It is presumably better at distinguishing between the two. The tool is trained and tested with newly assembled and curated dataset. The authors claimed to have achieved drastic reduction in misclassification as compared to SignalP-HMM (lower false positive but higher false negative rates). (http://phobius.sbc.su.se/)
PrediSi (Hiller *et al.*, 2004)	PWMs	Gram+, Gram-, Euk	This Java-based prediction tool uses three matrices (Euk: [-16, +4], Gram+:[-21, +1] and Gram-:[-16, +2]). Data is extracted from UniProtKB/Swiss-Prot Release 42.9 with a total number of 2,783 eukaryotic, 236 Gram+ and 557 Gram- sequences. By using a normalized score of between [0, 1], it allows for comparison between the different matrices. It achieves notably better accuracy for the Gram- dataset as compared to the Gram+ and Euk data when it is benchmarked against SignalP (HMM and ANN versions). (http://www.predisi.de/)

RPSP (Plewczynski *et al.*, 2008)	ANN	Gram+, Gram-, Euk	This method uses two ANN with feed-forward, multi-layer architecture and back-propagation learning algorithm. The combined ANN is more accurate than either the ANN solely trained for eukaryotes or prokaryotes. It claims to be capable of rapidly distinguishing SP from non-SP with high accuracy. The accuracy of the identification of cleavage sites is around 73-78%. Dataset is extracted from Swiss-Prot Release 49.4. Only sequences with amino acid at position -1 that appear in these sets: Euk (A,C,G,L,P,Q,S,T) and Bac (A,G,S,T) are included. (http://rpsp.bioinfo.pl/)
SigCleave (Rice *et al.*, 2000)	PWMs	Gram+, Gram-, Euk	One of the simplest approaches used for the prediction of SP cleavage sites. It uses the modified method for the treatment of positions -3 and -1 in the matrix (von Heijne, 1986). Two weight matrices are constructed for the positions from -13 to +2: (a) prokaryotes (based on 36 aligned sequences) and (b) eukaryotes (based on 161 aligned sequences). Originally developed by Peter Rice in 1989, it has since been modified by Alex Bleasby. It is available as part of the EMBOSS package. (http://emboss.sourceforge.net/apps/cvs/emboss/apps/sigcleave.html)
SigHMM (Zhang and Wood, 2003)	HMM	Human, Mouse	This method uses the popular HMMER package version 2.2 (Eddy, 1998) to generate profile HMMs to model the tri-partite regions in SPs following a previous method (Nielsen and Krogh, 1998). Training data is from human while testing data is from mouse; both sets originate from Swiss-Prot Release 40. The method was later updated using HMMER version 2.3.2 and tested with experimentally verified SP datasets (Zhang and Henzel, 2004). (http://share.gene.com/zhang.wood.bioinformatics.2003/sighmm/index.html)
SignalP (Nielsen *et al.*, 1997; Nielsen and Krogh, 1998; Bendtsen *et al.*, 2004b)	ANN	Gram+, Gram-, Euk	The most popular tool for SP prediction. Version 1.0 and 3.0 are based on ANN. Version 3.0 uses the same architecture as Version 1.0 except that the model has been retrained. It uses two networks to recognize windows containing cleavage sites from non-cleavage sites (*C-score*) and another to distinguish windows with SP and non-SP ones (*S-score*). The maximal combined score termed *Y-score* is used to identify the cleavage site. The *S-score* was subsequently replaced by *D-score* in Version 3.0, which is average of mean *S-score* and the maximal *Y-score*. Different window sizes are used in encoding the ANN.

	HMM		The accuracy of version 2.0 may not be as good as its ANN version, however, this version is better at detecting the presence of SPs and discriminating between SPs and uncleaved signal anchors. (http://www.cbs.dtu.dk/services/SignalP/)
Signal-BLAST (Frank and Sippl, 2008)	Pairwise alignment	Gram+, Gram-, Euk	The pairwise local alignment search tool, BLASTP (Altschul *et al.*, 1997) lies at the heart of this approach. Input sequence is queried against two curated datasets simultaneously to determine to which it is likeliest to belong to. The datasets essentially consist of SPs- and non-SPs- containing sets and a "signal peptide bias" is used to calibrate the comparison. This tool should be easier to maintain compared to other approaches. (http://sigpep.services.came.sbg.ac.at/signalblast.html)
Signal-CF (Chou and Shen, 2007)	OET-KNN + Scaled Window/ Subsite coupling/ Fusing	Gram+, Gram-, Euk	This tool consists of a two-layer predictor where a query protein is first identified as secretory or non-secretory (OET-KNN as classifier) before determining its cleavage site if it is a secretory protein by capitalizing on the subsite coupling effects of {-3, -1, +1} along a protein sequence and fuses the results derived from many width-different scaled windows through a voting system to determine the cleavage site. This tool is better at identifying SP cleavage sites and non-secretory of bacterial sequences as evident from its benchmark against SignalP (HMM and ANN versions) and PrediSi using Swiss-Prot Release 50.7. (http://www.csbio.sjtu.edu.cn/bioinf/Signal-CF/)
Signal-3L (Shen and Chou, 2007)	Similar to Signal-CF	Gram+, Gram-, Human, Plant, Animal, Euk	This tool expands from the original second layer of Signal-CF to two layers thus creating a three-layer predictor to achieve improvement in accuracy. Data used is from Swiss-Prot Release 50.7. (http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/)

SIG-Pred (Bradford, 2001)	PWMs	Gram+, Gram-, Euk	Developed as part of a thesis work. The tools uses the approach from (von Heijne, 1986a) but substituted the matrices with the updated ones derived from 1,011 eukaryotic sequences, 266 Gram- and 141 Gram+ sequences which is obtained from Swiss-Prot Release 29 (Nielsen *et al.*, 1997). Good accuracy achieved for Gram- data, moderate for Gram+ data but poor results for eukaryotic data. (http://bmbpcu36.leeds.ac.uk/prot_analysis/Signal.html)
SOSUIsignal (Gomi *et al.*, 2004)	Indices (include hydrophobicity	Pro, Euk	A tri-module system where the first module recognizes the hydrophobic segment in the 100 residues at the N-terminus. The second module determines if a sequence possess a SP or otherwise by using a *SS-score*. The final module discriminates SPs from signal anchors using a *SP-score*. Datasets are extracted from Swiss-Prot Release 40.
SPEPlip (Fariselli *et al.*, 2003)	ANN + PROSITE pattern	Gram+, Gram-, Euk	The training data from (Menne *et al.*, 2000) is used to develop this predictor that uses two ANN: (i) *netC* for identifying the cleavage site; (ii) *netS* for the detection of SPs. PROSITE pattern PS00013 is used to discriminate between lipoproteins and SP-containing chains. (http://gpcr.biocomp.unibo.it/cgi/predictors/spep/pred_spepcgi.cgi)
SPOCTOPUS (Viklund *et al.*, 2008)	ANN+ HMM	No division	An extension of the OCTOPUS tool (originally used for transmembrane protein topology prediction) to provide combined prediction of SPs and membrane protein topology. The training data is the compiled dataset from (Käll *et al.*, 2004). (http://octopus.cbr.su.se/index.php)

than other approaches especially in light of new data. For earlier reviews on machine learning techniques, refer to (Ladunga, 2000) and (Schneider and Fechner, 2004).

These prediction methods have been applied with varying degree of success in scores of studies, including the large-scale Secreted Protein Discovery Initiative (SPDI) which sought to discover novel human secretory and transmembrane proteins (Clark *et al.*, 2003); the genomic analysis of SARS-associated *Tor2 isolate* coronavirus (Marra *et al.*, 2003); identification of secreted proteins in bacterial proteomes (Bendtsen *et al.*, 2005a) and parasitic nematodes (Elling *et al.*, 2009; Nagaraj *et al.*, 2008). Likewise, tools such as SignalP (Bendtsen *et al.*, 2004b) have been used to annotate database sequence entries in which experimental evidence is lacking. These tools can be useful for locating homologous sequences or predicting the correct start codon as well, since SPs are situated at the N-terminus of proteins (Nielsen and Krogh, 1998).

6.2 Motivations

Several existing tools such as SigCleave (Rice, 2000) were built upon earlier matrices (von Heijne, 1986) which were in turn sampled from a much smaller aligned sequences. These matrices will need to be updated to reflect the correct observations of the relative frequencies of the residues as many sequences have since been generated. Thus, it follows that matrix-based tools recorded a drop in accuracy when they were tested with much larger and recent test sets. For instance, SigCleave registered an accuracy of 52% (Menne *et al.*, 2000), lower than an earlier claim. When it was later updated with a modified weight matrix for the positions nearby the cleavage site, its accuracy remains around 54.7% (Zhang and Henzel, 2004). The need to update is similarly applicable to machine learning based methods (or "active learning"). Their system parameters will have to be re-optimized and the underlying predictive models will have to be rebuilt particularly if the new sequences introduce distribution that is largely different from the existing. In such situation, some models may fail since a key assumption is made that the underlying data distribution is supposed to be similar/same.

In addition, the present matrix-based approaches almost entirely employ the window frame of [-13, +2] which was first established by (von Heijne, 1986). It was later affirmed that these positions were sufficient to achieve maximal accuracy for the prediction of SP cleavage sites (Chou *et al.*, 2001). However, the results from this study seem to suggest that there may be room for improvement (*Chapter 4*). Then, there are several approaches that rely on fixed-size window that do not address short SPs whose lengths fall within the length; their datasets omit such sequences from consideration.

Furthermore, there were two benchmark studies by (Menne *et al.*, 2000) and (Zhang and Henzel, 2004), that specifically compared the SP prediction tools available at that time but a number of newer tools have since been introduced with supposedly faster or more accurate prediction (Table 5). Also, majority of the comparison studies were conducted during the development of their respective prediction tool (Table 5) with several studies that involved only a subset of sequences or tools (Klee and Ellis, 2005) or non-experimentally verified SPs (Bagos *et al.*, 2008). In some cases, the performance indicators reported actually differ in the aspects that were being investigated (e.g. discrimination of SP or non-SP proteins OR/AND identification of the clevage site) (Gomi *et al.*, 2004). Thus, it will be useful to examine these tools simultaneously to allow proper comparison.

The majority of existing techniques exhibit high accuracy in distinguishing SP- from non-SP-containing sequences, but fare moderately in identifying SP cleavage site. It was even reported that a dismal one-third rate of inaccuracy was found in many of the putatively assigned cleavage sites (Zhang and Henzel, 2004). Hence, it is likely that existing tools have not been able to fully capture the essential information to develop a robust predictive method.

6.3 Methodology

In this work, two objectives are set:

(i) To develop predictive method that is able to *detect* the presence of SP and *identify* its cleavage site. The following sections describe a novel approach that is developed based upon the consolidated insights obtained in the earlier studies on the sequence (*Chapter 4*) and structure related (*Chapter 5*) to SPs;

(ii) To conduct a benchmark study on the existing prediction tools (Table 5) and the newly developed methods using cleansed datasets.

6.3.1 Preliminary testing using position weight matrices (PWMs)

The aim of this preliminary test is to evaluate the predictive results of using positions flanking the cleavage site compared to existing approaches (Table 5) that used mainly positions from the SP region. Some of the methods even incorporated positions well into the *n-region* of SPs.

Here, the position weight matrices (PWMs) described in Table 4 form the basis of the approach for testing. As the simplest type of probabilistic pattern method, PWM or also known as position-specific scoring matrix or sometimes known as profile (though profiles are technically different as they are more complicated and allows for gaps), is an ungapped table that records the relative frequency of amino acid residues at different positions that are observed within a fixed-size window frame.

Sequences are aligned with respect to the cleavage site and analyzed for their amino acid composition. Amino acid physico-chemical properties (Table 4) are excluded due to the poor accuracy upon our initial investigation. Multiple putative motifs are identified and the corresponding PWM is constructed to capture the patterns observed in the aligned motifs. The matrix score $S_{i,c}$ of an amino acid residue i at column c is calculated using the equation:.

$$S_{i,c} = -\log_2 \left[\frac{n_{i,c} + b_{i,c}}{(N_c + B_c) fb_i} \right] \qquad (1)$$

where $n_{i,c}$ and $b_{i,c}$ are the observed counts and pseudocounts (Henikoff and Henikoff, 1996) respectively; Nc is the total number of observed sequences and B_c is the total pseudocounts introduced to reduce the distortion due to the size of the training set and it is assigned the value of $\sqrt{N}$ (Lawrence *et al.*, 1993). $b_{i,c}$ is estimated by:

$$b_{i,c} = B_c fb_i \qquad (2)$$

where the background frequency, fb_{i} for each residue i is estimated from the frequency of occurrence of that residue in all the sequence positions outside of the calculated motif block. Thus, to obtain the score for a sequence fragment of length w, the score for each residue found in this sequence fragment is added.

One challenge in developing PWM is to determine the boundary of the sequences block that will be used for constructing the matrix, or in other words, finding the optimal window size for each organism group. Different matrices are therefore required as the distribution of amino acids differs among the organism groups (Euk, Gram+ and Gram-) (Table 4).

In all three matrices (Table 4), the occurrence of the hydrophobic residues Leu dropped markedly between P7 and P5. These hydrophobic residues regularly serve as the demarcation between the *h-region* and *c-region*. In addition, helix-breaking residues such as Pro and Gly (commonly occurring at positions -6 to -4) are taken into consideration since the earlier modeled SPase I-substrate complex depicts (i) a pronounced backbone twist between P3 and P1'; (ii) a beta-conformation in the c-region (*Chapter 5*).

Hence, different window sizes around the positions just described, are tested ([-6, +4], [-6, +3], [-6, +2], [-6, +1], [-5, +4], [-5, +3], [-5, +2] and [-5, +1]) using five-fold cross validation (see *section 6.4.2* for description on cross-validation). The optimal PWM is selected based on the highest cross-validation rate for each organism group. The input sequences are subsequently scored using a sliding window scheme where the corresponding PWM for each organism is successively aligned to every position from the N- to C- terminal direction (Figure 18).

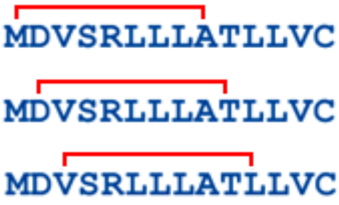


Figure 18: Diagrammatic representation of a sliding window scheme. A window of fixed-size is matched to the sequence in succession. Each of the matched sequence fragment is scored based on the matrix scores tabulated in Table 4.

The weight/score of each aligned residue in an aligned window is added to yield an alignment strength score (S_m) and the alignment/window with the highest S_m above a pre-determined threshold is considered as the likeliest to contain a cleavage site. The results of using the PWMs in this test are shown in later sections for ease of comparison and illustration.

6.3.2 Development of a sequence-structure SVM approach

Drawing on the previous findings (*Chapter 5*), the aim is to exploit the spatial constraints and the structural conformations of the SP in binding with the SPase I. This approach borrows from the concept that highly conserved residues often imply functional importance and they commonly appear at important sites in the protein 3D structure. Furthermore, it is known that a 3D structure of a protein is much more stable relative to the sequence in terms of divergence where even distant relatives within a protein family exhibit same overall topology and architecture (Panchenko and Bryant, 2002). Therefore, it can be reasoned that this method possibly can identify and characterize the functional sites by aligning the motifs flanking the cleavage site even if sequence similarity appears to be low. This could be useful in tackling the sequence variability, a notoriety that is closely associated with SPs.

Relating this concept to the bounding partner of SP — SPase I (*Chapter 2*), apart from the similarities and differences between the organism groups, in general, although there is limited sequence identity between the SPases I for all three organism groups, there are several critical regions (including the catalytic domain) which are homologous and they are located close to the SPase active site (Paetzel *et al.*, 2000). Additionally, a recent study has shown that the structure of SPase I does not change substantially upon substrate binding, thus suggesting changes which is locally confined (Musial-Siwek *et al.*, 2008). Also, examination of the structural pockets and grooves of SPase I-substrate complex (*Chapter 5*) reveals that these compartments can only accommodate certain types of amino acid residues constrained by their size. This preference is observed in Table 4 that shows the occurrences of specific residue at certain position (*Chapter 4*). Thus, it will be interesting to capture limited region around the cleavage site for all three organism groups and model them to investigate their differences.

In this approach, which we named as "SNIPn", information from the 3D structure is represented as feature vectors to capture the spatial constraints into our models. To generate the feature vectors, the previous structural model of *E. coli* SPase-I-SP complex (*Chapter 5*) is used as a template, specifically only the SP ligand portion ("template"). The fragment sequence to be modeled ("target") is aligned to the template and input to the homology modeling software called MODELLER (Sali and Blundell, 1993). Five models that mimic closely the structure of the template are generated. These models are optimized using the variable target function method with conjugate gradients and then refined using molecular dynamics with simulated annealing (Sali and Blundell, 1993). A statistical potential method called the "Discrete Optimized Protein Energy" (DOPE) (Shen and Sali, 2006) is used to guide the assessment of the modeled structures in which the one with the lowest DOPE score is considered as the likeliest model for the given alignment and template. The 3D coordinates (X, Y and Z axes) of each amino acid residue (represented by the backbone atoms N, C_α, C and O) of this optimal model are subsequently extracted. An additional feature called thermal factor or B-factor (Rhodes, 2006) that indicates the relative mobility of that particular atom extracted from the respective PDB file is also included. Together, these features are encoded thus resulting in a total of 13 features (3 x 4 + 1) per residue and 143 features per sequence for the window of 11 residues. However, due to the much expansive feature space compared to the data that is currently available to train the model to an appropriate level, we have added an additional features by encoding the linear sequence located within the [-6, +5] frame using binary encoding (+1 for presence and -1 otherwise) where each amino acid residue is represented by a vector of 20 features e.g. Ala is represented by (**1**,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1), Cys (-1,**1**,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1) and so on thus resulting in another 200 features (20 x 10). To construct the models based on these descriptors, Support vector machines (SVM) is selected for its powerful classification abilities as evident from its successful applications in many biological problems (reviewed in Zheng, 2004).

SVM is a statistical learning method based on the structural risk minimization principle that can handle linear as well as non-linear data (Burges, 1998). It is well suited to perform classification and complex pattern recognition tasks. In particular, it performs efficiently with high dimensional inputs. Additionally, SVM ability to outperform other machine learning techniques in the absence of large training dataset is attributed to its excellent generalization in dealing with unseen data (Zheng, 2004). Several groups have developed SVM-based approaches for SP prediction (Cai *et al.*, 2003; Mukherjee and Mukherjee, 2002; Sun and Wang, 2008; Vert, 2002; Wang *et al.*, 2005), but none have approached the problem from the structural angle.

An outline is given here since detailed explanation on the use of SVM for pattern recognition has been described in other literature (Burges, 1998; Joachims, 2002; Vapnik, 1998). The gist of SVM is essentially to (i) map input feature vectors to a high dimensional feature space through a mapping function Φ in conjunction with a kernel *K* which measures the similarity between different members of the dataset; (ii) construct an optimal hyperplane in the new feature space. Hyperplanes connect the bounds on the true error and separate the examples into positives and negatives.

The optimal separating hyperplane maximizes the margin of distance between the hyperplanes hence uniquely classifying the data into positive and negative examples.

Given a sample set *S* of *n* examples,

$$(x_1, y_1),\ldots,(x_n, y_n) \text{ where } x_i, \in R^n, y_1 \in \{-1,+1\} \tag{3}$$

There is at least one hyperplane that can separate the sample into positive examples at one side of the hyperplane and negative examples at the other side with a weight vector *w* and threshold *b*. This is given by the function

$$h(x) = sign(w \bullet x_i) + b \text{ where } \begin{cases} +1, \text{ if } (w \bullet x_i) + b > 0 \\ \text{-1, else} \end{cases} \tag{4}$$

for each example (x_1, y_1). If there are more than one hyperplanes, SVM selects the one with the largest margin Δd – the distance from the hyperplane to the closest training examples.

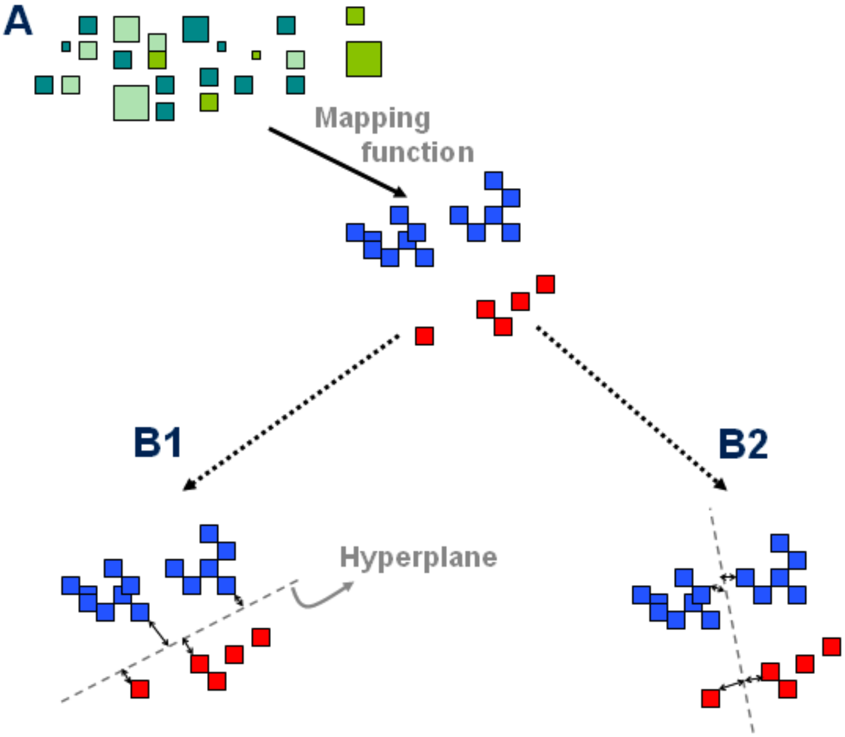


Figure 19: (A) Raw datasets are transformed to feature vectors and mapped to a higher dimensional feature space. (B1) and (B2) depict the possible scenarios where the examples can be separated using different hyperplanes.

Figure 19 shows an example where there are two hyperplanes that can separate the training set as illustrated by the two scenarios B1 and B2. There is only one hyperplane with maximum margin for every separable training set. Hence, the task is to find this optimal hyperplane. In the context of this work, the learning task is to classify a given sequence fragment as containing a cleavage site or otherwise. A decision value is then assigned for each predicted sequence fragment.

6.4 Training and Testing

6.4.1 Preparation of training data

The training and validation sets consist of 2,352 experimentally verified SPs taken from SPdb 5.1 (*Chapter 4*). The dataset is further divided into eukaryotes (1,877 sequences), Gram+ (168) and Gram- (307) bacteria. Only the first 70aa residues are retained as the datasets (basis of using this length has been described in *Chapter 2*). The diverse number of sequences should provide a good statistical sampling of the sequences that are likely to be found with the given motif.

Using matrix [-6, +4] (window size=10) as an illustration, six amino acid residues before the cleavage site and four residues after the cleavage site are used to generate the positive training dataset for the preliminary test using our PWMs. All the other windows that do not overlap exactly with these positions are used to generate the negative training set, including segments such as [-7, +3], [-8, +2] … [-5, +5], [-4, +6] and so on.

For our new method — SNIPn, positions from -6 to +5 are used to generate the positive training set as these positions are found within the groove where the binding occurs. Any positions outside of this region constitute the negative training data. Additionally, non-secretory sequences are also partitioned using the same window size of 11 to generate the negative instances. A summary of the training set is given in Table 6. All the attributes in the datasets are linearly scaled to the range [-1, 1] prior to training and testing. Scaling helps to avoid attributes with greater numeric ranges from dominating those of smaller range. It also reduces numerical calculation difficulties (Hsu *et al.*, 2008). The same method of scaling is applied to input data that require prediction.

Table 6: Training datasets that are used for the PWM preliminary test and development of SNIPn. Non-secretory sequences are omitted due to the availability of large negative instances. * only the first 11 residues from the MP portion is used to achieve a trade-off between computation time and performance.

	Positives	**Negatives**		**TOTAL**
		Outside of Window	Non-secretory	
Euk	1,877	42,313*	-	44,190
Gram+	168	9,903	10,080	20,151
Gram-	307	18,106	18,420	36,833
TOTAL	2,352	70,322	28,500	101,174

6.4.2 Parameter selections

To minimize the probability of overfitting, cross-validation procedure is applied where the training data is divided into two, consisting of training and testing sets. This process of partitioning the dataset is repeated until the testing set (1/N of the size of the entire dataset) is cycled through the entire dataset, exactly N times for an N-fold cross-validation (Figure 20). This procedure ensures that each testing set is predicted exactly once. Further, to attain a balance representation of positive and negative instances in the partitioned dataset particularly in the presence of greater number of negative examples, we stratify the dataset such that roughly similar ratio of positive and negative examples is maintained across the different partitioned sets. This procedure is applied to both our preliminary testing of using PWMs and to SNIPn.

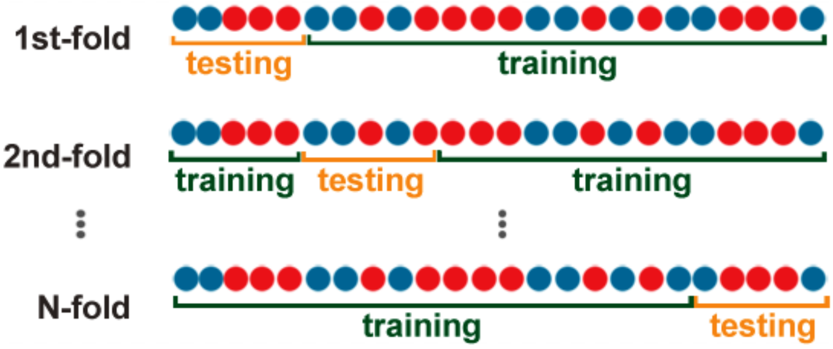


Figure 20: Schematic representation of cross-validation with positive (blue circle) and negative (red circle) instances scattered through the datasets. A non-overlapped testing set is sampled through each fold.

The Gaussian radial basis function (RBF) (given by the equation $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ where $\gamma > 0$) is selected as the kernel function for SNIPn as RBF has been shown to (i) handle non-linear relation between class labels and attributes well; (ii) use less hyperparameters which ultimately affects the complexity of model selection (iii) has less numerical difficulties (Chang and Lin, 2001; Hsu *et al.*, 2008). To determine the optimal parameter pair (*C*, γ) of the RBF kernel for the respective model of each organism group, the parameters are subjected to grid search using 5-fold cross-validation. *C* is the penalty cost for misclassification while γ controls the degree of nonlinearity of the model. Once the optimal parameter pair (rendering highest cross-validation rate) is determined, the entire training set is retrained to generate the final model for classification.

6.4.3 Testing and evaluation

Dataset preparation (filtering and redundancy reduction)

The test sets used for the benchmarking are generated following the "*SP Filtering Rules*" (*Appendix B*) with some adaptations (Table 7):

(i) The positive set consists of 270 secreted recombinant human proteins taken from (http://share.gene.com/cleavagesite/index.html) (Zhang and Henzel, 2004). As the original study did not create any negative dataset to test the specificity of the tools, we have to separately create a negative set using the 270 human non-secretory proteins from the dataset used for the construction of SigHMM (Zhang and Wood, 2003);

(ii) Using the dataset described in *Chapter 4*, there are 2,352 positive instances (Euk:1,877; Gram+:168; Gram-:307) and covers most of the data used to develop the majority of the prediction methods compared here. The positive set is matched by an equal number of negative instances for each organism group. The negative dataset is a mix of cytoplasmic and nuclear proteins (applicable to eukaryotes only). Proteins from other subcellular localizations are excluded since it is difficult to state unequivocally whether they are secreted (Bendtsen *et al.*, 2004b). Similarly, single-pass type II membrane proteins that contain signal anchor are not used as well since the majority of the entries are predicted (http://www.expasy.org/cgi-bin/lists?annbioch.txt) (labeled "Potential"). We use the "KW" field, instead of "SUBCELLULAR LOCATION" phrase under the "CC" field, to locate the cellular localization due to its more succinct description. Organellar proteins and proteins containing cTP or mTP are also removed. Additionally, entries with the keyword "Secreted" appearing under the "KW" field are removed (e.g. F13A_HUMAN which is cytoplasmic in most tissues, but it is secreted in the blood plasma as well). Finally, visual inspection is conducted to remove atypical sequences e.g. ATX8_HUMAN which consists of only Ms and Qs in its sequence. In the positive set, unlike other studies, we do not exclude sequences with SPs that are shorter/longer than the average since such sequences do exist, and they have been annotated and verified. Omitting them is synonymous to fitting data to model instead of the reverse.

(iii) A new dataset is extracted from Swiss-Prot Release 57.0 following the '*SP Filtering Rules'* (*Appendix B*). Sequences (both positive and negative) which are present in (ii) are deliberately omitted (based on their Swiss-Prot *ID* and *accession number*) from this dataset to create a new dataset that is novel for the majority of the tools (except those that have been recently updated such as Signal-BLAST). This would minimize any prior advantage enjoyed by the tools in predicting SPs from sequences similar to those "seen" before. Manual inspection of the filtered data reveals that many of the entries are putative. Only those cleavage sites that are highly probable based on the evidence from literature are retained otherwise more than 90% of the bacterial entries and more than 50% of the entries in eukaryotes would have been eliminated had the "*SP Filtering Rules*" being applied.

The test sets are maintained in equal balance between the positive and negative instances to ensure there will be no bias in the assessment of the tools. Duplicates are removed from the positive datasets while negative datasets (non-secretory proteins) are further reduced using CD-HIT (version 3.1.1) (Li and Godzik, 2006) to create a diverse set of sequences. Whenever possible (either bounded by the minimal number of sequences for testing or the lowest CD-HIT threshold that can be set), the lowest possible threshold is adopted.

Exclusion of previous datasets

The popular training/testing sets (Nielsen *et al.*, 1997; Menne *et al.*, 2000) are not adopted in this evaluation since they are derived from earlier Swiss-Prot releases (Release 27.0 and Release 38.0 respectively). The second dataset (SPdb 5.1 which is derived from Swiss-Prot Release 55.0) used in this study are inclusive of these sequences.

Omission of prediction tools

SPEPlip (Fariselli *et al.*, 2003) is omitted due to the lack of facility for large-scale testing. A number of methods that are unavailable for testing are omitted as well. They include several neural network-based approaches (Jagla and Schuchhardt, 2000; Li *et al.*, 2008; Reczko *et al.*, 2002); SVM-based approaches (Cai *et al.*, 2003; Mukherjee and Mukherjee, 2002; Sun and Wang, 2008; Vert, 2002; Wang *et al.*, 2005); a profile HMM-based method called CJ-SPHMM (Chen *et al.*, 2003); matrix-based approach that uses the concept of information theory (Liu *et al.*, 2005); a BLOMAP-encoding scheme to transform input sequences (Maetschke *et al.*, 2005); a hybrid approach that uses bio-basis function neural networks and decision trees (Sidhu and Yang, 2006); a global alignment approach based on the Needleman-Wunsch algorithm (Liu *et al.*, 2007; Needleman and Wunsch, 1970).

Other tools such as *iPSORT* (Bannai *et al.*, 2002), *ProteinProwler* (Hawkins and Bodén, 2006) that are mainly used in subcellular localizations and N-terminus targeting signals (e.g. *Predotar* (Small *et al.*, 2004)) prediction are omitted as well since they predict the presence of SPs but do not indicate the cleavage sites. We have also omitted specialized tools such as *SecretomeP* which predict non-classical SPs i.e. signal sequence remained uncleaved (Bendtsen *et al.*, 2005b).

Setup of prediction tools

For PWMs testing, three PWMs with size of *W*x20 each, are derived from the aligned motif block of width *W* for the different organism groups where $W_{\text{euk}} = W_{\text{Gram+}} = W_{\text{Gram-}} = 10$ based on the motif/matrix [-6, +4]. The thresholds or cut-offs that achieve the maximal accuracy are selected (Euk:5.65; Gram+:6.68; Gram-:5.10) (Figure 22).

On the other hand, SNIPn uses a window size of 11 for all three organism groups in its prediction. Figure 21 shows the system architecture for our SVM-based classifier where the LIBSVM package (version 2.8.8) (Chang and Lin, 2001) is used as the SVM implementation. Modifications are made to the program code to output the decision values instead of the label. The optimal parameter pairs (C, γ) of the RBF kernels for the respective organisms are empirically determined based on 5-fold cross-validation (Euk(1, 0.05); Gram+(2, 0.005); Gram-(1, 0.01)). The *–w* option is adjusted to account for the imbalanced dataset (more negative instances than positives at a ratio of 119:1 for bacteria and 22.5:1 for eukaryotes) to avoid model overfitting.

For *PrediSi*, the web server is used instead of the standalone version due to the discrepancy in their results. The standalone version reported numerous inaccurate predictions even for the same input sequence. The prediction results are converted to 0 if the result field “Signal Peptide ?” indicates an “N” otherwise the predicted cleavage site is recorded if a “Y” is shown.

For tools that employ different models/matrices for different organism group (e.g. *SignalP*, *Signal-CF* etc.), the corresponding matrix is selected accordingly. *Signal-3L*, in particular, allows for six selections: (i) human; (ii) plant; (iii) animal; (iv) Gram-positive; (v) Gram-negative; (vi) "other-eukaryotic". The authors' categorization method as shown in (*Online Supporting Information B*: http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/Data.htm) is used to classify and select accordingly the right matrix for a given input sequence.

For *SigCleave*, the default threshold (*-minweight*) of 3.5 is used to filter the results as suggested in their documentation. For *SigHMM*, a returned score below -5 (Zhang and Wood, 2003) is deemed to indicate a non-secretory protein, otherwise the cleavage site is reported since the sequence is considered as a secretory protein.

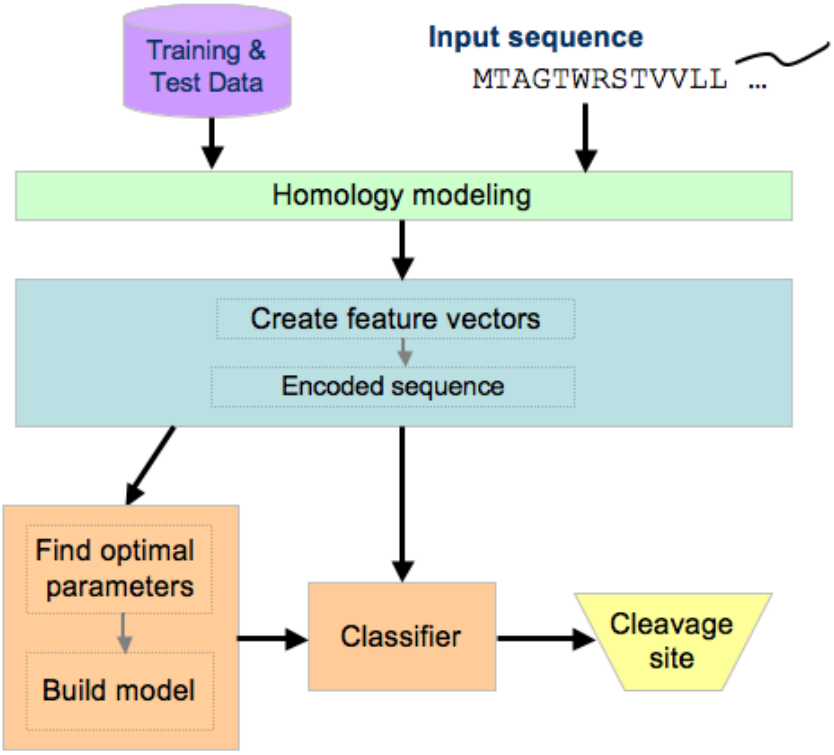


Figure 21: The architecture of our SVM-based prediction system — *SNIPn*. Sequences (either from the user or the training/testing datasets) are first encoded to create the feature vector representing the sequence. The encoded feature vector is sent for classification task. The predictive model used in the classifier is the optimal model selected during the training and testing phases.

For *Philius*, all its predicted values were subtracted by one from its originally predicted values except when the value is already zero. It is highly possible that there is a bug in *Philius* since the returned value is always one extra position away, i.e. instead of 24aa, it predicts 25aa. This bug has been reported to the authors.

For Signal-Blast, the detection mode is set to "SP4 - Only Detect Cleavage Site". For all other tools not specifically mentioned, we have used their default system settings with no additional parameter changes made except selecting the corresponding organism matrices, when available. All parameters for each tool are maintained the same in all three experiments and the experiments are carried out on 32-bit Intel-based desktop computers equipped with 2GB of memory. It should be noted that running on 64-bit machines generates different results during the structure modeling phase and machine learning phase due to the higher precision available for floating point numbers.

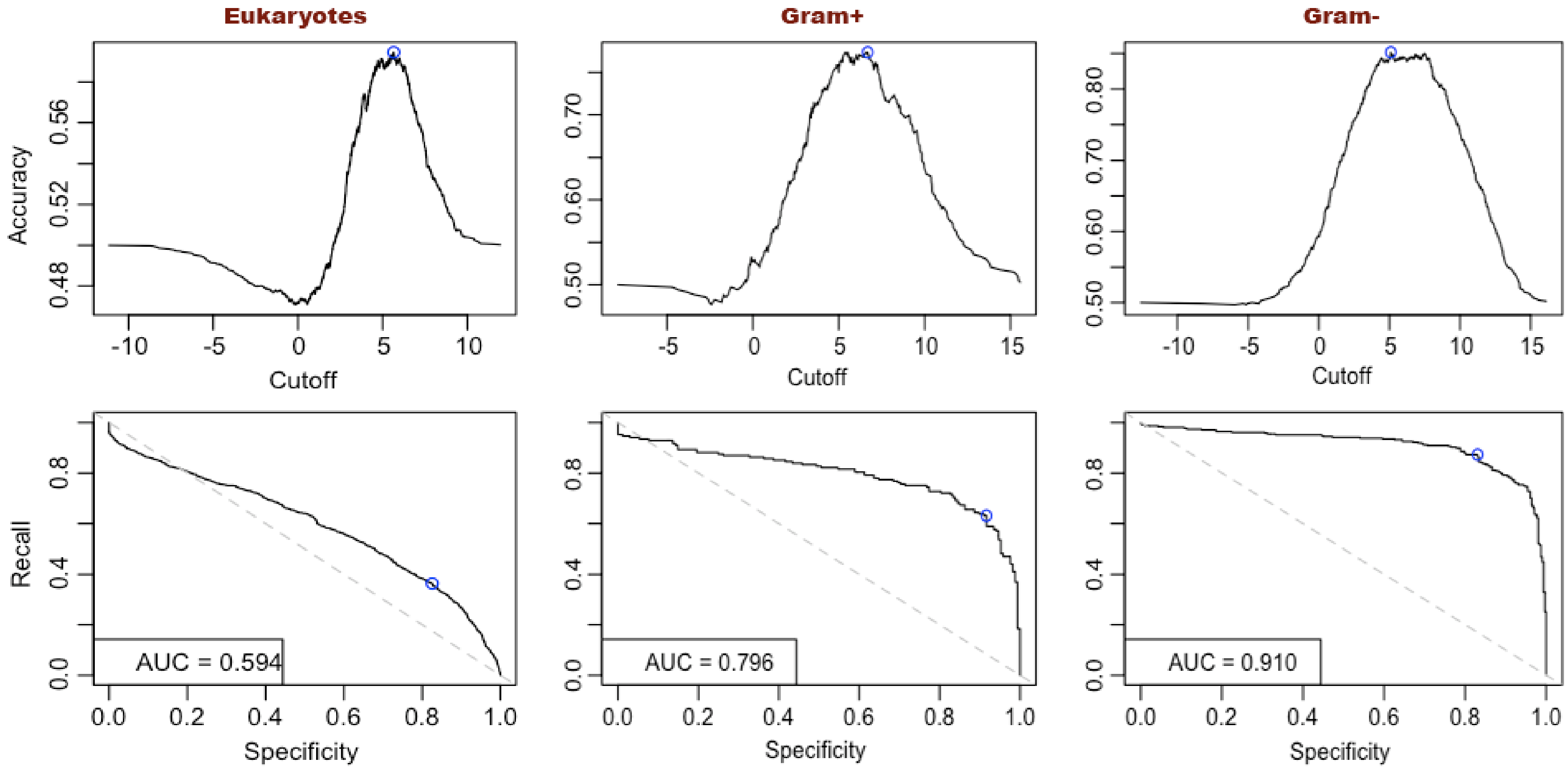


Figure 22: The charts in the first row plot the accuracy against the varying cut-offs for the three organism groups. The second row shows the corresponding ROC curves. The (blue) circle located in each chart denotes the selected threshold that yields the maximal accuracy. The charts are generated using the R statistical package (R Development Core Team, 2009) augmented with two additional modules: the ROCR (Sing *et al.*, 2005) and Brendano's dlanalysis (http://github.com/brendano/dlanalysis/tree/master).

Evaluation of prediction tools

All results from the different tools are standardized to the following:

$$\text{Results} = \begin{cases} 0\text{, if predicted as non-secretory protein OR unable to predict the position} \\ \text{position of cleavage site, if predicted as secretory protein} \end{cases}$$

To evaluate the predictive performance of all the prediction tools, we compute *sensitivity* (Sn), *specificity* (Spc), *accuracy* (Acc) and Matthews' Correlation Coefficient (MCC) (Matthews, 1975). The equations are given by:

$$Sensitivity(Sn) = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity(Spc) = \frac{TN}{TN + FP} \tag{6}$$

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{8}$$

where *Sn* and *Spc* measure the fraction of positive instances and fraction of negative instances respectively which have been correctly predicted. *Acc* computes the fraction of positive and negatives instances predicted correctly. *Mcc* returns a value that is between 1 (perfect prediction) and -1 (inverse prediction); the value zero denotes a random prediction. Briefly, sequences that possess cleavable SPs that are subsequently predicted with the correct cleavage sites are designated as true positives (TP). Those that are predicted with the wrong cleavage sites are treated as false negatives (FN). Conversely, sequences without cleavable SPs that are predicted with one are classified as false positives (FP) whereas predictions specifying an absence of SP are considered as true negatives (TN).

6.5 Results

The results from the three experiments are shown in Table 8 and Figure 24(A)-(H). Figure 24(A) depicts the overall accuracy values for each method across the three experiments. Experiment 2 and 3 provide values for three organism groups while Experiment 1 essentially measures the results for eukaryotes alone. Most tools attain accuracies that are well over 80%, consistent with what have been reported in many earlier studies, which were without complete details of specificity and sensitivity. A breakdown of the prediction results measured by sensitivity and specificity for each experiment, give us a better account of the strength and weakness of each tool.

6.5.1 Results from Experiment 1

The first experiment uses 270 eukaryotic (human) sequences with experimentally verified SPs, from the study by (Zhang and Henzel, 2000).

Based on the results from this experiment (Figure 24(B) and Table 8), Signal-BLAST predicts the highest number of correct positive instances (i.e. best sensitivity) (97.8%). This is dramatically reversed when it is scores 81.5% in specificity upon tested with negative instances in which it is tasked to distinguish between secretory and non-secretory proteins. This is attributed to the need for Signal-BLAST, which uses a pairwise alignment, to find a delicate balance between the two types of datasets in order to achieve a good discrimination. SignalP scores the second best accuracy with the ANN version (87.2%) marginally outperforming the HMM version (85.6%).

Signal-CF and Signal3L which adopt the “subsite-coupled model” achieve accuracies of 77.4% and 81.3% respectively. The results are lower than those reported in the authors’ publications using the same dataset. Manual inspection of Signal-3L revealed that there was a mistake quoted in their publication (Shen and Chou, 2007). For the entry (Swiss-Prot ID:Q6UXL0), the cleavage site was reported as 28aa instead of the correct 29aa that the authors indicated in their supplied supplementary data ("*Online Supporting Information B: Signal-CF dataset–supp-B.txt*"). From our examination, Signal-CF and Signal-3L identify the cleavage site at 63aa and 28aa respectively based on the input sequence of length 70aa. When we reduced its evaluation length to match the length reported in their publication (MQTFTMVLEEIWTSLFMWFFYALIPCLLTDEVAILPAPQNLSVLSTNMKHLL MWSPVIA), Signal-CF and Signal-3L reported SPs of 29aa and 28aa. Furthermore, we noted that selecting the correct species option in Signal-3L is critical; otherwise a markedly different length of SP is reported. Signal-CF, on the other hand, is extremely sensitive to the different lengths.

Among the tools compared, SOSUIsignal, SPOCTOPUS and our *PWMs* method rank lowest in sensitivity (18.9%, 39.3% and 26.7% respectively). This is likely because the identification of cleavage site was not their priority. SOSUIsignal was developed to discriminate SPs from non-SPs sequences, while SPOCTOPUS was developed as a combined predictor for SPs and membrane protein topology. For our PWMs, it is likely that the necessary information may not be adequately found within the limited window size. On the other hand, SNIPn returns moderate results although the sensitivity may actually be lower had the model been adjusted to increase the specificity.

Other methods generally return accuracies that are above 80%. However, closer inspection reveals that while the specificity values are impressive, their sensitivity values are largely in the moderate range of 70% to 79%.

Table 7: Description of the three datasets developed for benchmarking the thirteen SP prediction tools, including ours. Only the first 70aa of the sequence are retained as input. Negative dataset are subjected to redundancy reduction. *T* denotes sequence identity threshold set for redundancy reduction. [1] From a first-pass-filtered set of 9,851 reduced to 4,989 upon redundancy reduction (*T*=40%) and atypical/spurious sequences removal before arriving at this filtered set; [2] From a first-pass-filtered set of 427 reduced to 230 (*T*=40%); [3] From a first-pass-filtered set of 370 reduced to 307 (*T*=65%); [4] From a first-pass-filtered set of 8,930 reduced to 4445 (*T*=40%); [5] From a first-pass-filtered set of 110 reduced to 61 (*T*=40%); [6] From a first-pass-filtered set of 290 reduced to 150 (*T*=40%).

	1: Zhang and Henzel, 2004 (Experimental data)	**2: Dataset used in this study** (Extracted from SPdb 5.1 which is in turn derived from Swiss-Prot Release 55.0)	**3: UniProtKB/Swiss-Prot Release 57.0** (excludes the dataset that we have used in this study)
Positive	270 human secreted recombinant proteins	2,352 secretory proteins consisting of: - Eukaryote: 1,877 - Gram+: 168 - Gram-: 307	228 secretory proteins consisting of: - Eukaryote: 199 - Gram+: 17 - Gram-: 12
Negative	270 human non-secretory proteins extracted from SigHMM (Zhang and Wood, 2003) dataset which is in turn derived from Swiss-Prot Release 40.0.	2352 non-secretory proteins - Eukaryote: 1,877 (Cytoplasmic: 939; Nuclear: 938) [1] - Gram+: 168 (all cytoplasmic) [2] - Gram-: 307 (all cytoplasmic) [3]	228 non-secretory proteins Eukaryote: 199 (Cytoplasmic: 100; Nuclear: 99) [4] - Gram+: 17 (all cytoplasmic) [5] - Gram-: 12 (all cytoplasmic) [6]

Table 8: Benchmark results of the thirteen prediction tools (*Table 5*) including ours, based on our three standardized datasets. Equation (5-8) are used to measure the predictive performance of these tools. (Abbreviations used: Sn=Sensitivity; Spc=Specificity; Acc=Accuracy; MCC=Matthews' Correlation Coefficient). [1] Used with HMMER 2.3.2 with cut-off score set at -5 (Zhang and Wood, 2003) and the updated model (Zhang and Henzel, 2004); [2] Version 3.0; [3] Authors updated system with UniProt 14.6 (Swiss-Prot Release 57.0); [4] Version 1.0.1. * Our methods

	1 : Zhang and Henzel, 2004 (Experimental data)				2 : Dataset used in this study (SPdb 5.1; derived from SwissProt Rel. 55.0)				3: Swiss-Prot Release 57.0 (excludes all dataset used #1 and #2)			
Methods	Sn	Spc	Acc	MCC	Sn	Spc	Acc	MCC	Sn	Spc	Acc	MCC
PWMs *	0.267	0.759	0.513	0.030	0.391	0.833	0.612	0.249	0.368	0.820	0.594	0.211
SNIPn *	0.900	0.819	0.859	0.721	0.994	0.802	0.898	0.811	0.408	0.737	0.572	0.153
Philius	0.704	0.952	0.828	0.677	0.742	0.968	0.855	0.729	0.728	0.961	0.844	0.708
Phobius	0.637	0.978	0.807	0.654	0.750	0.982	0.866	0.752	0.711	0.987	0.849	0.726
PrediSi	0.726	0.974	0.850	0.723	0.769	0.986	0.878	0.774	0.750	0.974	0.862	0.742
RPSP	0.730	0.989	0.859	0.744	0.806	0.996	0.901	0.816	0.794	1.000	0.897	0.811
SigCleave	0.541	0.878	0.709	0.445	0.612	0.824	0.718	0.446	0.618	0.860	0.739	0.493
SigHMM[1]	0.707	0.937	0.822	0.662	0.561	0.963	0.762	0.572	0.596	0.952	0.774	0.587
SignalP[2] ANN	0.785	0.959	0.872	0.756	0.856	0.965	0.911	0.826	0.842	0.987	0.914	0.838
SignalP[2] HMM	0.759	0.952	0.856	0.725	0.832	0.974	0.903	0.814	0.833	0.969	0.901	0.810
Signal-BLAST[3]	0.978	0.815	0.896	0.803	0.881	0.809	0.845	0.692	0.825	0.794	0.809	0.619
Signal-CF	0.648	0.900	0.774	0.566	0.768	0.905	0.837	0.679	0.750	0.890	0.820	0.647
Signal-3L	0.737	0.889	0.813	0.633	0.787	0.920	0.853	0.713	0.715	0.934	0.825	0.665
SOSUIsignal	0.189	0.926	0.557	0.170	0.232	0.925	0.578	0.217	0.232	0.921	0.577	0.212
SPOCTOCUS[4]	0.393	0.907	0.650	0.350	0.503	0.889	0.703	0.442	0.408	0.899	0.654	0.352

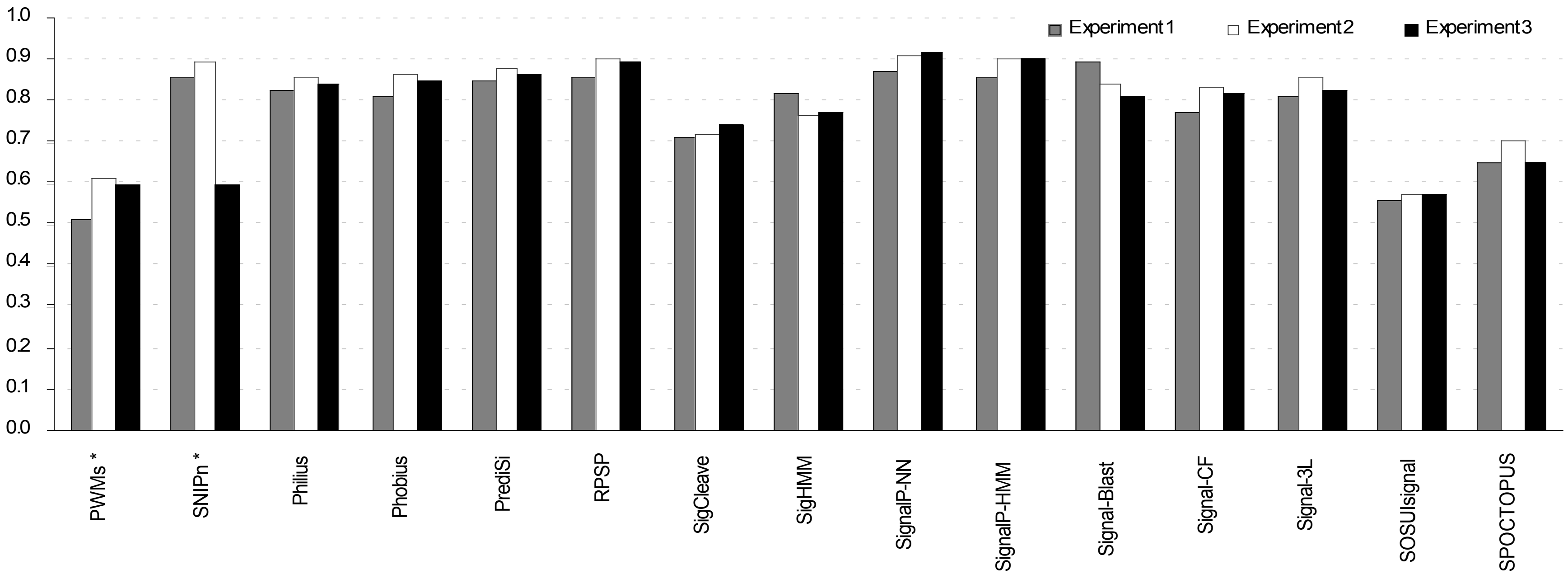
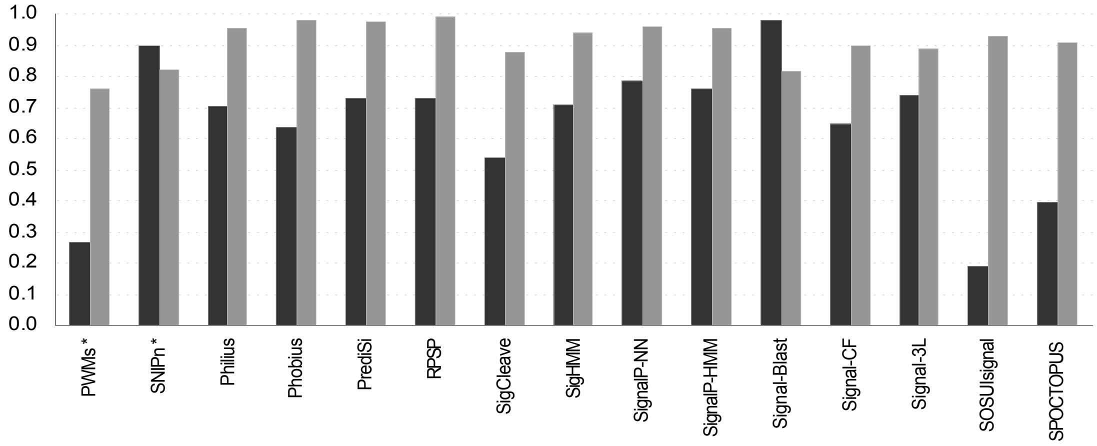
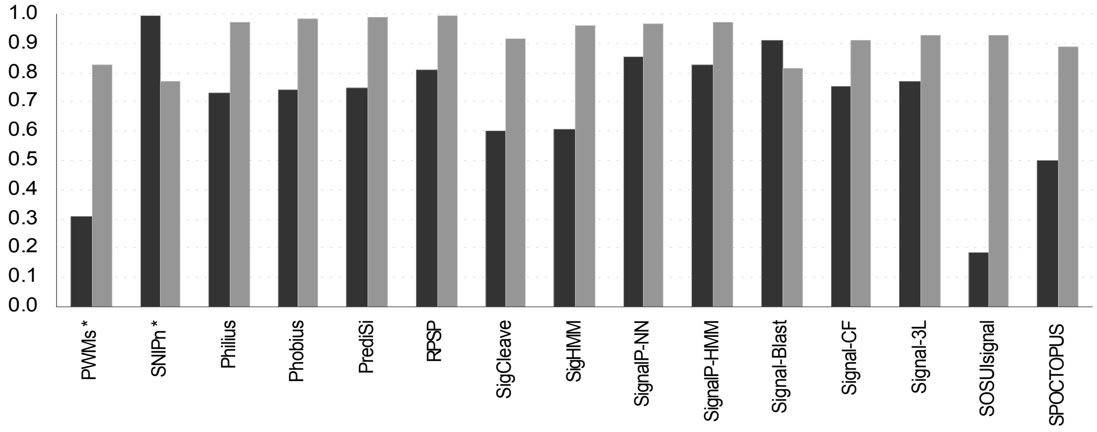


Figure 23: Aggregated results from all three experiments. Accuracy results from all three experiments are provided here. For each tool, there are three bars, representing each experiment (gray bar: Experiment 1; white bar: Experiment 2; black bar: Experiment 3). * denotes the methods that we have developed and tested in this study.

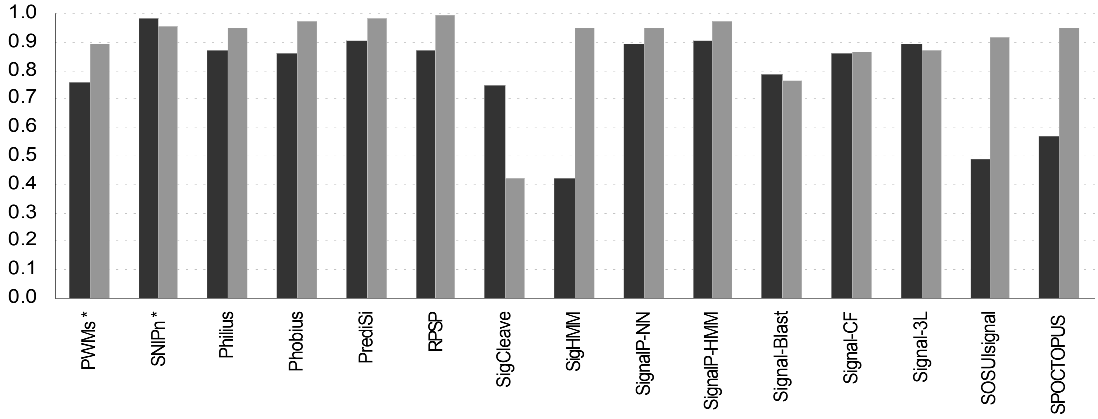
(A) Experiment 1: Eukaryotes (human)



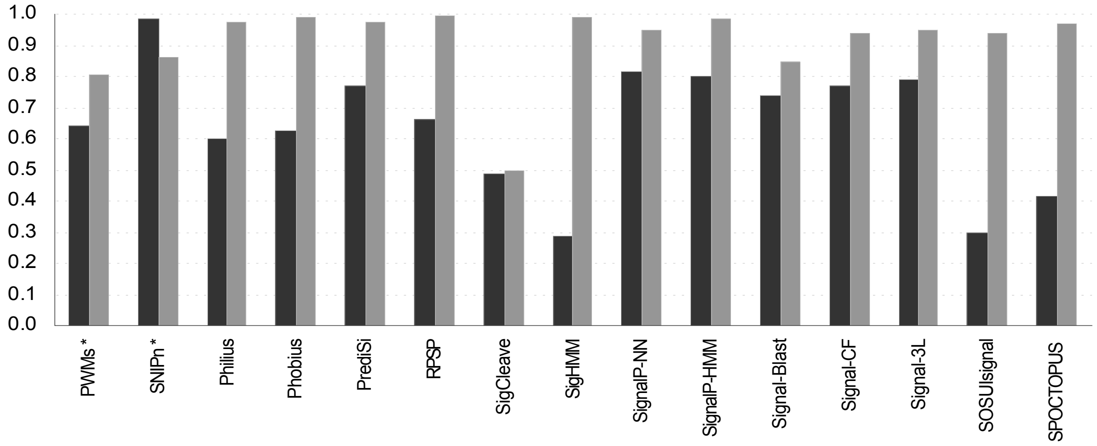
(B) Experiment 2: Eukaryotes

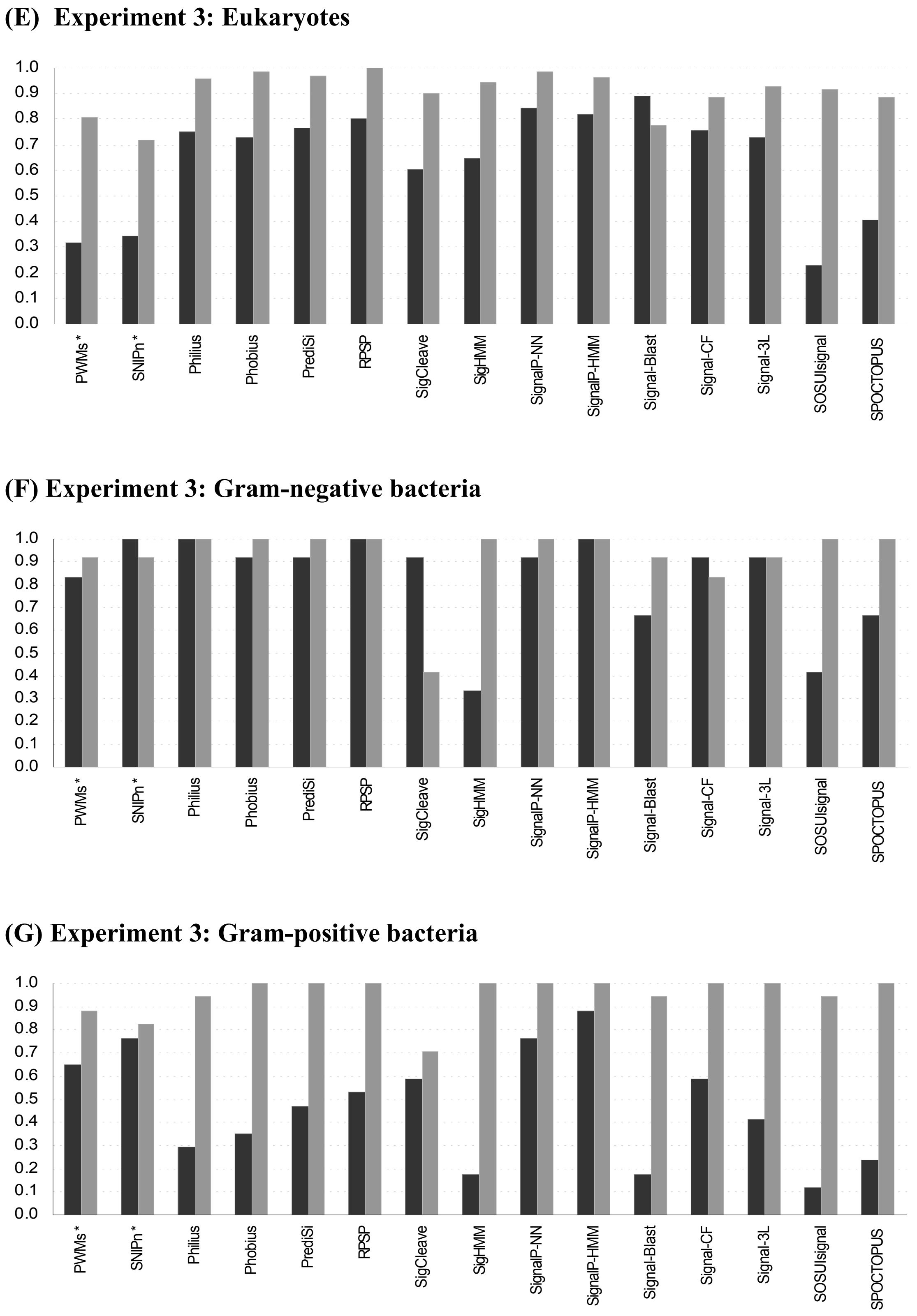


(C) Experiment 2: Gram-negative bacteria



(D) Experiment 2: Gram-positive bacteria





(E) Experiment 3: Eukaryotes

(F) Experiment 3: Gram-negative bacteria

(G) Experiment 3: Gram-positive bacteria

Figure 24: (A) Experiment 1 involves eukaryotic (human) sequences only; (B)-(D) Results from Experiment 2 separated into the three organism groups: eukaryotes, Gram+ and Gram-bacteria; (E)-(G) Results from Experiment 3 separated into the three organism groups. The bars colored in light gray represent the specificity while the darker bars represent the sensitivity of the predictive tools.

6.5.2 Results from Experiment 2

This experiment recruits a much larger dataset consisting of 4,704 sequences that are spilt into positive and negative datasets of equal size. The negative set consists of a mix of cytoplasmic and nuclear sequences in eukaryotes. The dataset is further divided into three organism groups (details available in Table 7).

Our *PWMs* achieves overall accuracy of 61.2%. Detailed inspection of the result breakdown (classified by organism groups) reveals that the *PWMs* obtain good results in the Gram- and Gram+ datasets but not in the eukaryote set. SNIPn achieves an overall accuracy of 89.8% where the breakdown of the accuracies is 88.4% (Euk), 97.1% (Gram-) and 92.3% (Gram+) respectively. The result from the bacteria group is better than the leading tool – SignalP-ANN (Gram-:92%; Gram+:88.1%) and SignalP-HMM (Gram-:93.8%; Gram+:89.0%) (refer to Figure 24 (D), (E), (G) and (H)).

Interestingly, *PWMs* outperforms SigCleave in both sets. For the Gram- set, *PWMs* reaches accuracy of 82.6% against SigCleave (58.5%) while for the Gram+ set, *PWMs* achieves accuracy of 72.3% against SigCleave (49.4%). The results of SigCleave are marginally lower than that of SigHMM (71.8% against 76.2%). When we examine their results further by looking at the individual data groups (Figure 24(C)-(E)), in particular within the bacterial datasets; SigHMM obtained sensitivity values of 42.0% (Gram-) and 28.6% (Gram+), respectively. A comparable drop in both measurements is observed in Experiment 3 (cf. next section). This is possibly attributed to the newer bacterial sequences that have become available since the model was constructed. SigCleave experiences a similar fall in performance for the Gram- (sensitivity:74.6%; specificity:42.3% and accuracy:58.5%) and Gram+ (sensitivity:48.8%; specificity:50.0% and accuracy:49.4%) datasets. Other prediction tools generally maintain similar trend as observed in the previous experiment, though their sensitivity values are considerably lower in the Gram+ bacteria dataset compared to the Gram- bacteria and eukaryote datasets.

Between Signal-CF and Signal-3L, it seems that the additional classification of sequences into specific groups (e.g. plant, human, animal etc.) used in the latter method do not seem to generate much advantage over the former approach, and may potentially lead to overfitting.

6.5.3 Results from Experiment 3

New datasets have been extracted from Swiss-Prot Release 57.0 (totaling 412,525 entries) in this experiment (details available from Table 7). This dataset represents a fresh challenge for majority of the tools except for Signal-BLAST which has been recently updated with Swiss-Prot Release 56.6. The results are presented in Table 8 and Figure 24 (F)-(H).

Here, SignalP (both ANN and HMM versions; with HMM scoring higher than ANN) again presents consistently high results. The sensitivity values for other tools plummet particularly when tested with the Gram+ dataset. This drop is particularly acute for Signal-BLAST, despite its recent update. We checked the distribution of the data but do not note any significant differences compared to the previous two datasets.

SNIPn experiences a considerable accuracy drop in the eukaryotic prediction (57.2%). It is probable that the model might have overfitted during the training which explains the good results observed in Experiment 2 since part of the data was used. However, there is also the possibility that the may not be adequate data to construct the model to a sufficient level. This is because the linear sequences have been mapped to a much bigger feature space using their 3D coordinates. When we tested the model by adding some of the new sequences directly into the existing model without any further optimization, the model is able to predict correctly for the sequences. Thus, more sequences will be required to determine the reason.

6.6 Discussion

6.6.1 Simple model or sophisticated model

It was previously suggested that non-linear feature may be involved in the recognition of cleavage site (Nielsen and Krogh, 1998), thus this perhaps helps to explain the better accuracy achieved by machine learning based techniques. However, in this study, it is observed that the performance gap between the more simplistic matrix-based approaches (e.g. PrediSi, SigCleave and our *PWMs*) and the sophisticated machine learning-based approach is not significantly wide. Given the appropriate selection of the window size, matrix-based approach can achieve competitive results.

Alignment-based technique such as Signal-BLAST, SigHMM can be tuned to be more sensitive in identifying cleavage site, but at the expense of its specificity or vice versa. For instance, when we submit the sequence from human carboxylesterase 2 isoform 1 (GenBank GI:37622885) to Signal-BLAST, a markedly different entry (Swiss-Prot ID:ICAM1_HUMAN; with reported cleavage site of 27aa) was returned as the top hit with an assigned cleavage site of 19aa. Such method generally may not be suitable for detecting sequences that share weak homology, since it is highly dependent on how the tool balances sensitivity with specificity. Thus, compared to matrix-based approach, these methods (in particular Signal-BLAST) will probably require more effort in updating it with new releases/updates from sequence databases such as UniProtKB-Swiss-Prot to remain relevant as long as newer sequences are deposited.

6.6.2 Larger dataset and window size

The majority of the prediction tools achieve better results for the eukaryotic datasets compared to the bacterial datasets. This is likely attributed to the larger data size that is available to build models that adequately describe the underlying distribution. Conversely, the results from our methods are the opposite. This is probably due to selection of the window size ([-6, +4] and [-6, +5] for *PWMs* and SNIPn respectively), which may not have been sufficient to capture the necessary information for the eukaryotic group, thus explaining the uniformly mediocre accuracies observed in the eukaryotic sets in all three experiments.

For the bacteria datasets, we have demonstrated that it is possible to achieve competitive results when compared against the counterparts even with the reduced number of positions. This stands in contrast to previous studies that have advocated for greater number of residues to be considered within the window frame to achieve maximal accuracy (Chou *et al.*, 2001). The results from the predictive methods here lend support to the earlier notion that cleavage recognition possibly do not require residues located significantly further upstream of the SP, as well as residues in the MP portion located far away from the cleavage site, at least this appear to be case for the bacteria sequences that have been tested in these experiments.

Interestingly, in an earlier separate experiment in which we constructed the SVM models for SNIPn using only the structural information as feature vectors (each residue represented by four atoms and the corresponding X, Y, Z coordinates), the top thirty-five most predictive features are measured and extracted using the following method:

$$F - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{9}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{10}$$

where recall is essentially the same as sensitivity (Equation 5). F-score (also F-measure or F1 score) is yet another way to measure the accuracy of performance (1 being the best and 0 at its worst).

Remarkably, the result (Figure 25) manifests similar graph/pattern as what have observed in the earlier sequence analysis study (*Chapter 4*) even though only one structure is employed in the homology modeling to generate the 3D features. The conservation of the residues at specific positions (e.g. the motif at -3, -1) is clearly observable. Furthermore, it is also observed that the distribution of feature vectors of eukaryotes clearly differs from bacteria even at the 3D structure level. It may well require positional information beyond what have been explored here (i.e. -6 to +5 or +4) which explain for the poor results for our *PWMs* and SNIPn methods when tested on eukaryotic sequences. The availability of eukaryotic structure data will help greatly in explaining these differences, particularly for SNIPn, where we have used Gram- 3D structure as the template to model for all three organism groups.

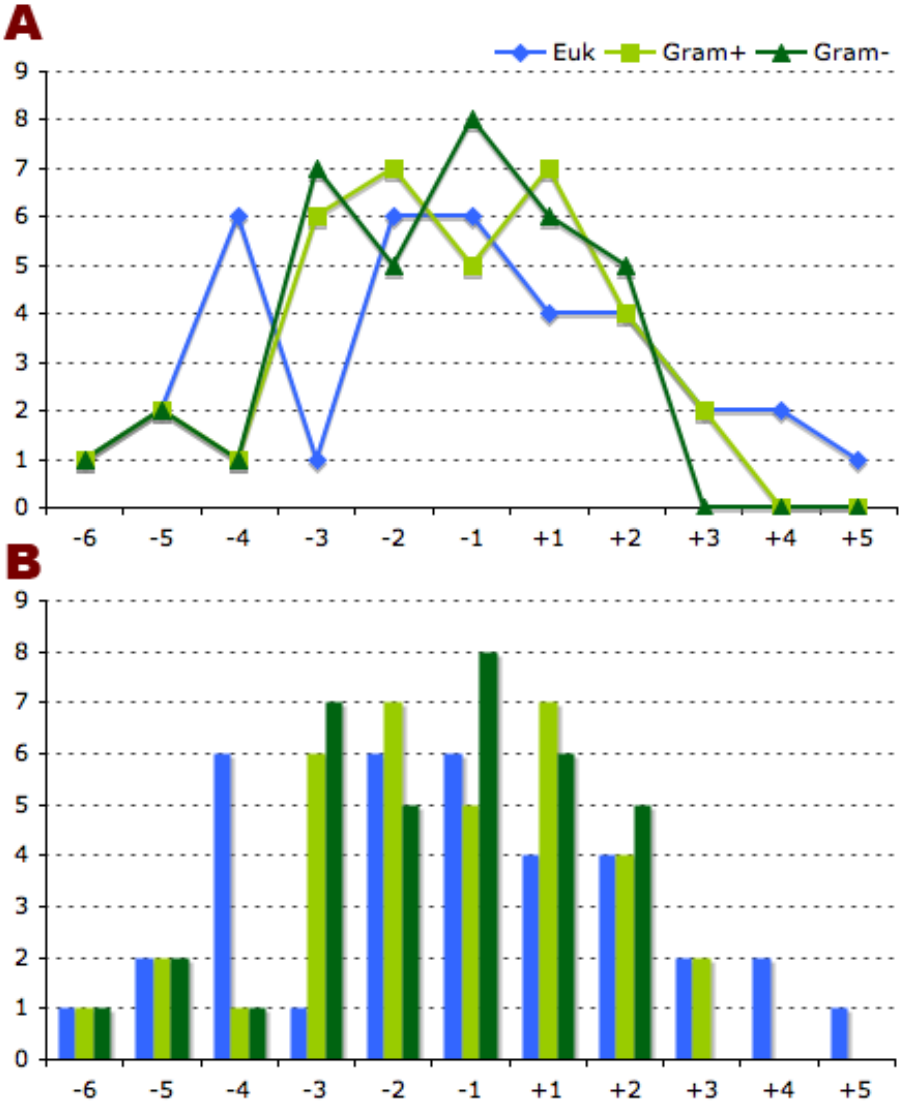


Figure 25: Top thirty-five attributes/features that are the most predictive or significative as measured according to F-score values through a five-fold cross-validation. The data is represented in two format (A) line graph and (B) bar chart. X-axis shows the positions within our employed window of [-6, +5] for the SVM-based approach. The junction -1/+1 denotes the SP cleavage site. Y-axis tracks the number of features that represent a residue at a particular position within the window of [-6, +5].

Except for the leading tool SignalP which have been rather successful in their prediction for all three organism groups across the three experiments, majority of the tools will probably require active learning or regular update to the underlying model to remain relevant. The consistency observed in SignalP (both ANN and HMM versions) may be attributed to its more complex models and robustness of its method where various scoring schemes are devised to tackle different aspects (including SPlikeness, the probability of a segment containing the cleavage site and so on). Also, the window frame employed are also relatively wider (Euk:[-11,+2], Gram-:[-21,+2], Gram+:[-15,+2]). However, it should also be noted that many of the sequences, particularly in Experiment 3, have been assigned their cleavage sites using SignalP, thus the data may contain certain biasness. More new data that have not been putatively assigned by computational tools will be needed to ascertain the true veracity of these tools.

6.6.3 Single-step or two-step prediction task

In general, most tools encounter little difficulty in distinguishing between secretory and non-secretory proteins. This is evident from the high specificity achieved even when they are tested with new datasets. Other studies involving discrimination between signal anchors and SPs reach similar conclusion (Nielsen and Krogh, 1998). Identification of the corresponding cleavage site clearly remains the challenge.

In contrast to the majority of the prediction methods where they divide the prediction problem into separate two tasks, namely (i) discrimination between secretory and non-secretory proteins; (ii) prediction of the cleavage site, we tackle both simultaneously without distinguish the tasks. Here, we have demonstrated that it is possible to pinpoint the correct cleavage site (at least in the bacteria group) and discriminate between SP- and non-SP-containing sequences with reasonable accuracy. However, separating into two tasks will enable better results since it allows for further optimization. This is demonstrated by SignalP where they have designed multiple scoring schemes and models to capture and measure the different aspects. Our preliminary results of separating the tasks have demonstrated better sensitivity.

6.6.4 Assessment of our method

Based on the preliminary test using our relatively smaller PWMs when compared to existing matrix-based methods including SigCleave, we have been able to achieve favorable results in predicting bacteria sequences even with reduced positions. However, the result is reversed in the case of eukaryotic sequences. Hence, it is likely that this group will require a much larger matrix size to achieve optimal prediction accuracy, at least for PWM-based method. The fact that using our smaller PWMs is able to deliver competitive results suggests that possible further exploration by focusing on flanking regions around the cleavage site.

Our approach is therefore to use homology modeling (one *E. coli* SPase I as template) to generate the theoretical binding models of SP and its receptor. The resulting structural information is extracted and modeled as feature vectors in SVM. The results have been encouraging, at least for the bacteria group even though the same result is not observable for the eukaryotes, which is explainable. One reason is the window size that is used for the eukaryotic group which may be inadequate to capture the necessary details. More importantly, our structural template for aligning the sequences is currently based on *E coli* alone and it is known that the accessibility to the active site and cleavage processing machinery differ for bacteria and eukaryotes, thus, the 3D conformation of the eukaryotic and bacterial SPs may again be different. Thus, replacing the *E. coli* template with the appropriate structure from eukaryote may actually offer a different result. Furthermore, due to the mapping of the features from a linear sequence to a 3D space, the hypothesis space has essentially increased tremendously. The current size of data therefore may not be sufficient to train the model to an appropriate level compared to the number of features that we have employed. When we tested using some of the new (eukaryotic) test sets by adding them to the SVM model (without any adjustment to the optimal parameters), the model is able to predict correctly for majority of the sequences. However, we are unable to ascertain this currently due to the unavailability of the eukaryotic structure.

One drawback with SNIPn is the intensive computation that is required to generate the 3D coordinates to be encoded as feature vector for the SVM. A possible remedy is to record sequence fragments that have undergone homology-modeling computation in a lookup table. Only non-existing ones upon consulting the table are computed and they are recorded into table for future lookup. Another solution is to deploy the program within a computational grid.

6.6.5 Testing of archaeal sequences

There are seven archaeal sequences that contain experimentally determined SPs in SPdb 5.1. The result from testing with this set of sequences using the three organism models is shown in Table 9. For SNIPn, the result from using the predictive model of Gram+ is the best (3 out of 7) followed by eukaryotes and Gram-. Similarly, in SignalP, the Gram+ model returns the best predictive results (6 out of 7). The results observed here is in agreement with the study (Bardy *et al.*, 2003) that archaeal SPs are more similar to the bacterial than the eukaryotic. Interestingly, the fact that the methods can predict the archaeal sequences using models from other organisms indicate their shared ancient origins, where they rely on these common SPs and translocation machinery to deliver their proteins.

In this mini study, SignalP achieves better results than SNIPn. When the detailed prediction values are inspected, we noted that our approach did manage to predict the correct site albeit with a weak score. This affirms the earlier discussion in the previous section (*single-step or two-step identification*) to split the prediction task into two sub-tasks to allow better optimization.

Table 9: Prediction results from SNIPn and SignalP (both ANN and HMM versions). Each row represent one entry/sequence extracted from Swiss-Prot which has been manually curated to possess experimentally determined SP. The first column (AR) lists the actual/known cleavage site while other columns tabulate the predicted values from each tool. GP, GN and EU represent the respective organism model that is used for the prediction (AR=Archaea; GP=Gram+; GN=Gram-; EU=Euk; HMM=Hidden Markov Model; ANN=Artificial neural networks).

SNIPn				**SignalP**					
				ANN	HMM	ANN	HMM	ANN	HMM
AR	**GP**	**GN**	**EU**	**GP**		**GN**		**EU**	
34	0	0	0	29	28	20	28	22	20
34	34	23	23	33	34	34	34	23	23
34	0	37	0	34	34	37	37	23	23
34	0	0	23	34	34	23	34	23	27
22	22	22	22	22	22	22	22	22	22
28	28	0	28	28	28	23	28	28	28
46	0	37	0	37	37	37	37	0	37

6.7 Summary

We have presented a novel method called SNIPn, which uses SVM to model sequence and structure information to achieve competitive accuracy in SP prediction. It offers as alternative way for SP prediction in bacteria sequences. Further availability of new data will help to improve the predictive models of SNIPn.

In this study, we have also evaluated thirteen of the most commonly used prediction tools that are available for testing. Majority of the tools are able to distinguish secretory and non-secretory proteins with little difficulty. The challenge clearly remains with pinpointing the correct SP cleavage site.

Although we have shown that it is possible to achieve comparable results without splitting the SP prediction task into two sub-tasks, doing so will likely lead to better results as the model for each task can be further optimized. The composite scoring schemes employed by SignalP essentially divide the prediction task into a number of separate steps, thus allowing each score to tackle a particular aspect of the prediction. Additionally, it can be observed that some methods are more susceptible to new changes to the datasets. These methods likely require regular updates to their underlying models to reflect the latest observations. Alignment-based and matrix-based methods are such examples, where the updates will allow proper tuning of their model parameters.

Chapter 7: Conclusion

7.1 Summary

This work began as a fascination over these short peptides that are found at the N-terminus of virtually every secreted and (some) membrane proteins. These peptides exist only transiently within the secretory pathway but they assume such diverse roles and exert multiple functions that have wide-ranging effects on all living organisms (*Chapter 2*). Recent findings are suggesting SPs of possessing a far more impressive functional repertoire than their sole targeting function. This has spurred renewed interest to elucidate their true functions. Already, several latest research works are revealing the prospect of more exciting discoveries that lie ahead, including the report of more effective inhibitor of bacterial SPase I (Buzder-Lantos *et al.*, 2009), the binding of SRP and SP catalyzes the RNA component of SRP to accelerate the interaction between SRP and SR (Bradshaw *et al.*, 2009) and the discussion of non-conventional secretion transport pathways (Nickel and Rabouille, 2009).

In an effort to contribute towards the understanding of SPs, in particular, the understanding of their substrate specificity and cleavage processing, in this work, we have developed a semi-automated pipeline through systematic approach (*Chapter 3*) to generate a SP-centric repository called SPdb (http://proline.bic.nus.edu.sg/spdb). SPdb has been carefully curated to remove inconsistencies and detectable errors. Entries with discrepancy between the literature and the database annotation were flagged. New error detection rules were devised and combined with existing best practices to form the "*SP Filtering Rules*" (*Appendix B*). This set of rules has greatly reduced the laborious effort to flag erroneous or inconsistent entries and standardized their removal. Although the resulting datasets have served as the foundation for this work, SPdb can be a valuable and useful resource for benchmarking and developing new prediction tools. Moreover, it can support further research into different scientific studies. It is also applicable to technological/industrial applications.

Since the publication of this work (Choo *et al.*, 2005), the Swiss-Prot team has resolved numerous erroneous entries that were identified in this work. The labels have been re-assigned and some of the more confusing practices (e.g. the different labels such as POTENTIAL, PUTATIVE to describe evidence support) have been improved. It is also encouraging to learn that references are now inserted directly next to the relevant field (in the latest UniProtKB), to indicate only the associated publications for that feature. This is what we have demonstrated with SPdb. Nevertheless, similar problems that have been described in this work (Table 2) remain observable in the new entries of UniProtKB/Swiss-Prot. As a result, it will be extremely difficult to extract the data directly from sources such as Swiss-Prot and work on it immediately without proper treatment of the entries, as already explained in *Chapter 3*. A similar pipeline as what we have proposed here might be needed to tackle those issues (Table 2).

Next, based on the cleansed datasets, we conducted a large-scale analysis of 2,352 experimentally verified SP-containing sequences involving prokaryotes and eukaryotes. When we measured the lengths of the different SPs for all the organisms, we observed large variations in the length distribution (Figure 8). This variance that resulted in largely different SPs, reportedly mediate/induce different rates of closure of the ribosome-translocon junction (Rutkowski *et al.*, 2001). Further classification of the organisms into eukaryotes, Gram+ and Gram- bacteria groups for analysis revealed several similarities as well as distinctive features.

From the analysis of these three groups of organisms in terms of their physico-chemical properties such as p*I*, aliphatic index, GRAVY score, hydrophobicity and net charge, we observed markedly different property values, even for those SPs which possess almost similar lengths (Figure 11). These variations are possibly employed as a means for the machinery to vary translocation efficiencies (Kim *et al.*, 2002) or as a "tuner" for modulation depending on conditions (Kang *et al.*, 2006). This could help explain the hyper-conservation of SPs that was observed in conotoxins (Olivera *et al.*, 1999). The toxins have to be targeted and localized rapidly and accurately. Hence, it is logical to maintain the conservation of SP sequences. Given these observations, our next step was to apply the same analyses as what we have performed on SPs to the MPs. The fact that SPs can vary to such an extent, suggests that they will probably require the coupled coordination of the MP counterpart in order to achieve such a feat.

Indeed, several studies have shown such "SP-MP coupling" between the two regions to produce an optimal pairing (Brockmeir *et al.*, 2006; Kim *et al.*, 2002). There were studies (Li *et al.*, 1988; Summers *et al.*, 1989; Summers and Knowles, 1989) that have demonstrated how the balance between the two regions could influence export efficiency. When we measured the net charge between the two regions, the SP region is visibly predisposed to positive net charge, particularly for the bacterial SPs (Figure 11). This observation was further examined in detail with the tabulation of the occurrences of amino acid residues at various positions (from P10 to P10', with P1/P1' denoting the cleavage site) flanking the demarcation between the two regions (Table 4). Charged residues were observed to occur more frequently at the MP, almost immediately upon the cleavage site.

We then examine the sequences for any observable sequence motif. Except for the higher incidence (or perhaps the lack) of certain types of residues at specific positions (Table 4), for instance, higher incidences of Leu upstream of P6 or P7, the most significant pattern remains the Ala-X-Ala motif (von Heijne, 1986a) which was the essence for the postulation of the "(-3,-1) rule". According to the rule, only small and aliphatic residues are allowed P3 and P1, and aromatic, charged, large polar residues including helix breaking residues such as Pro are prohibited. Nevertheless, there were small quantities of such prohibited residues occurring at those positions (Table 4). Moreover, this motif is observed to occur in only half of the total dataset for Gram+ (61.9%), Gram- (77.5%) and eukaryotes (61.6%). The lack of strong motif in eukaryotic sequences can possibly be attributed to the more complex mechanisms and structures required. Eukaryotic SPases I are known to be more complex than the prokaryotic counterpars (Paetzel *et al.*, 2002b). Regardless of this difference between the organism groups, P3 and P1 have been known to be critical recognition sites for SPases I (Karla *et al.*, 2005). This suggests that there is possibly non-canonical cleavage motif where other secretion pathway(s) is/are utilized to secrete these proteins.

In earlier chapter, we posed several questions related to the substrate specificity. Specifically, what are the determinants that ensure the high fidelity of SPase I excision to occur exactly after the Ala-X-Ala motif and not elsewhere? For non-canonical cleavage sites (i.e. non Ala-X-Ala preceded cleavage sites) that do not bear any sequence pattern, what are the factors that govern their identification? SPs have been reported with numerous roles and functions apart from its usual targeting function (*Chapter 2*). What is the recipe for encoding such enormous amount of information within the short peptide length without escalating complexity further? Using information theory, our investigation on the sequences did not yield further signs that were not already observable from the patterns along the sequence (Table 4;

Figure 10). Furthermore, how do the components of the machinery in the secretion pathway cope with the degenerate feature of SPs while simultaneously maintaining the high specificity and high fidelity requirements in the targeting, recognition and cleavage of SPs? Also, why do certain alterations (substitutions, insertions or deletions) or mutations are tolerated with muted or no effects while others lead to drastic changes which can lead to dire consequences (Gierasch, 1989)? We hypothesize that the answer may perhaps lies beyond the linear sequence.

It is known that a protein is relatively much more stable at the structural level during the process of evolution. It usually encounters little change to its shape/fold even though the sequence may undergo substantial changes (Eidhammer *et al.*, 2004). This essentially means that there may be muted effect upon single/multiple substitutions of the amino acid residues, while concurrently susceptible to (drastic) rearrangement in conformational structure for a particular change in another residue (Pidasheva *et al.*, 2005; Ronald *et al.*, 2008). Extending this concept further, we asked if this could account for the recognition of cleavage sites that do not conform to the distinctive Ala-X-Ala motif, including the non-canonical ones?

These questions led us to investigate the structure of SP in complex with its cleavage enzyme, SPase I. However, such a structure is not readily available. There are only four crystal structures (PDB IDs: 3IIQ, 1T7D, 1KN9 and 1B12) of *E. coli* SPase I in complex with other substrates such as inhibitor or lipopeptide that have been resolved through X-ray diffraction and archived in PDB as of this writing. Hence, through threading and homology modeling techniques, we have created a working model of *E. coli* periplasmic disulfide-bond A oxidoreductase (DsbA) 13-25 in complex with its endogenous SPase I (*Chapter 5*). The resulting model was found to be in good agreement with the known experimental data, thus supporting the validity of our model.

Based on our model, we have identified thirteen subsites (S7 to S6’) within the SPase I substrate-binding site which were found to have significant interaction with the substrate. At these sites, ten positions for hydrogen bonding were identified to possess high affinity binding, thus suggesting that the contact points between the enzyme and substrate may extend throughout the P7 to P6’ of the substrate. This observation is also supported by our survey of the amino acid residues surrounding the cleavage-processing site. These flanking residues are very likely to influence the cleavage processing and contribute to non-canonical cleavage sites that were observed in our earlier analysis of their sequences (*Chapter 4*).

In addition, we found that the subsites S3’/S4’ were able to alter their electrostatic requirements by varying their side chain conformations. This may help explain the propensity to interact with substrates with charged residues at these positions. Furthermore, the large cavity at S3’/S4’ subsites allows for the accommodation of medium and large residues. A pronounced twist was observed in the backbone between P3 and P1’ (Figure 14), which contain the cleavage site.

By combining the insights that we have gained from the sequence and structural analyses, we were motivated to apply the concept of the structure conservation discussed earlier, by modeling the cleavage site recognition problem through machine learning technique. The idea is to exploit the spatial features and constraints present in the SP-SPase complex. These features were extracted and used to train an SVM model. The resulting model was tasked to distinguish between secretory and non-secretory sequences, and also to pinpoint the cleavage site if a protein is identified as being secretory using SVM. The results from our predictive method called SNIPn, have been encouraging as demonstrated in our benchmark study (*Chapter 6*). It achieved accuracy that is competitive with existing state-of-the-art prediction methods in the bacterial datasets. Interestingly, our statistical examination of some of the most predictive attributes (Figure 25) in the trained model revealed similar positional patterns as manifested in the sequence analysis earlier (*Chapter 4*). A recent study has also suggested that so long as certain conformation to certain physical properties (charge, hydrophobicity etc.) or structural properties is fulfilled (Guo *et al.*, 2008), SPs will be functional as opposed to sole compliance to sequence conservation. An earlier study has also advocated for such overall and minimal requirements at the sequence and structural levels (Duffaud and Inouye, 1988). The need for structural conformation may well explain for the disruptive effects that were observed when charged or helix-breaking residues are introduced into the SPs (Oliver, 1985; Yamamoto *et al.*, 1989). It can also help explain the plasticity of eukaryotic and prokaryotic SPase I in mutual recognition of SP cleavage sites (Allet *et al.*, 1997; Osborne and Silhavy, 1993; Watts *et al.*, 1983)

With this combined use of sequence and structural features, SNIPn have been able to predict to certain degree of success (without additional modifications to the default parameters) for archaeal and viruses sequences (data not shown) as well. These results provide support for the shared evolutionary origin of these different organism groups. The availability of more sequence and structure data from these respective organisms can help to improve SNIPn predictive models, and this aspect of evolution could be further explored. Additionally, our structure-based approach could be useful in predicting proteins that do not possess ‘classical’ SPs such as FGF1, *Engrailed* homeoproteins and interleukin1 (Joliot *et al.*, 1998; Bendtsen *et al.*, 2005b). These proteins do not possess any characteristic motif, and they are secreted through various non-classical pathways (Prudovsky *et al.*, 2003). A recent study has shown that methods using amino acid composition, secondary structure and disordered regions could identify such proteins (Bendtsen *et al.*, 2005b).

Our predictive model is among the few tools to involve considerable portion of the MP. As described earlier, there may be an optimal pairing between a SP and its respective MP (Brockmeier *et al.*, 2006). Additionally, there has been an increasing body of evidence that suggests SPs are perhaps not as interchangeable as previously thought and they are likely not to be functionally equivalent (Kim *et al.*, 2002). Thus, if there is indeed such a coupling between the two segments (part of MP), it is more so that this relationship be admitted into the predictive models.

As to the extent of MP portion that is involved in the cleavage processing, this remains an on-going debate. Kajava *et al.* and several other research groups (*Chapter 2*) have advocated MP moiety for consideration that are located much farther downstream of the cleavage junction than what we have considered here. We reason that involving residues further downstream unnecessarily complicates the recognition process particularly when this process is only one of the many critical events that occur in the secretory pathway. The biological knowledge that we have gathered thus far also do not seem to point to this direction of the extensive involvement of the MP moiety (*Chapter 2*). Furthermore, the support for our stand also draws from the statistical results of the top most predictive attributes (*Chapter 6*) as well as the sequence analysis (*Chapter 4*). These results do not indicate any significant patterns beyond P5'. More importantly, we have also demonstrated through the practical implementation of a tool that exhibits good accuracy in the recognition of cleavage site through a limited number of positions upstream and downstream of the cleavage site, in addition to the superb discrimination of secretory and non-secretory signal sequences. Further availability of crystal structures and data relating to SPs shall provide clarification on this issue.

As of this writing, the unanswered questions pertaining to SPs far outnumber the answered. Despite more than thirty years since its discovery and the wealth of sequence data that has become available, it is remarkable that such a short signal sequence remains an enigma to scientists. Fortunately, and with tremendous anticipation, the improvements in technology are bringing us closer to critical understanding of SP and its underlying mechanisms. New tools and new methods will need to be devised to attain the enlightenment. The hasten adoption of computational methods complementary to traditional experimental approach shall produce a new synergy for us to revisit some of the assumptions that we have made herein as well as those that have been reported in current literature. It is all these unknowns that will bring about new exciting discovery and elucidation on SPs to harness them for potential use in drug design and industrial applications.

7.2 Key Contributions

This thesis makes several important contributions to the field of SP and related areas. They are summarized as follows:

- Creation of the largest and manually curated N-terminal SPs catalogue which is stored in SPdb relational database with integrated information derived from Swiss-Prot and EMBL sequence databases. The database is accessible from: (http://proline.bic.nus.edu.sg/spdb). Facilities to search and download are provided. The update process of SPdb is handled by a semi-automated pipeline. This ensures that the database can cope with the growing sequences, thus addressing one of the issues facing many existing databases where the dataset becomes outdated as time passes (Choo *et al.*, 2005). More importantly, SPdb serves as a useful resource to support scientific studies and methods development. It is currently in use by the global scientific community and it has been listed as a reference database under the Wikipedia (http://en.wikipedia.org/wiki/Signal_peptide)

- Formulation of new techniques and incorporation of several existing techniques for the detection of erroneous annotations and the removal of the affected sequence entries. This set of filtering rules is collectively known as "*SP Filtering Rules*" (*Appendix B*). Following the rules ensure a filtered set of SP sequences with vastly reduced errors

- Conducted a large-scale analysis of N-terminal SPs involving 2352 manually curated SP-containing sequences to study the physico-chemical properties and their composition. The result from the analysis are used in the development of our prediction method

- Development of a 3D computational model of *E. coli* SP in complex with its endogenous type I SPase using existing X-ray crystallographic data of *E. coli* substrate-SPase complexes. This work represents the few reports on the modeling of a substrate into the entire SPase I binding site, previous studies (Ekici *et al.*, 2007; Karla *et al.*, 2005; Paetzel *et al.*, 2000; Paetzel *et al.*, 2002a; Paetzel *et al.*, 2004). Romesberg lab of the Scripps Research Institute has requested the theoretical model for their design work in finding inhibitors of bacterial SPase I. Such theoretical model can serve as template for further investigation into antibiotic design

- Development of three PWMs, one for each organism group. These matrices require lesser number of positions than previously suggested positions to achieve favorable results in the identification of (bacterial) SP cleavage sites
- Development of a novel technique using SVM for SP prediction (presence detection and cleavage site identification). The resulting system called SNIPn, achieved accuracy that is competitive with existing leading prediction tools (in bacterial datasets where structure data is available) when it was benchmarked using various test sets (including new sequence data). While existing approaches have mainly explored the linear sequence and a few approaches that exploit the secondary information using physico-chemical properties, the combination of structure and sequence information through the use of homology modeling represents a fresh approach in SP prediction

- Conducted a comprehensive benchmark study involving all the leading SP prediction tools using standardized curated datasets to allow proper comparison between the different tools for the different organism groups. The last time such study was conducted by Menne *et al.* in 2000 but many prediction tools have since been introduced

- Errors were discovered in some of the publicly available resources during the course of this study and they were reported to the respective sources:
  - Sequences – the annotation errors were reported to Swiss-Prot
  - Prediction tools – the errors discovered while testing with the

prediction tool Philius were highlighted to one of the authors (Sheila M. Reynolds, University of Washington) where they have responded with modification to their prediction application

7.3 Future Direction

This project has considerable scope for expansion. The work done in *Chapter 3* paves the way for further development that will facilitate the extraction of other types of targeting signals and integrate them into SPdb to form a unified repository for all targeting signals. Facing the similar problem as SPs, these targeting signals currently do not have a dedicated repository that catalogue and curate them as what has been attempted in this study. Such central repository is crucial in providing a standardized resource for researchers and tool developers alike to benchmark their methods. Another line of work that is to explore the extraction of additional annotations such as mutation information on SP and other residue information related to SP through text-mining approach. This work could potentially help to discover new knowledge embedded within the voluminous literature. Preliminary work on this line of work is currently underway and part of the work has been published (Kanagasabai, 2007). Further potential uses based on SPdb have been outlined earlier.

The work done in *Chapter 4* has shown that the distribution and composition of amino acids and other their physico-chemical properties are markedly different in the different regions. In subcellular localizations, many works have exploited such features to develop localizations prediction tools. The analysis from *Chapter 4* similarly can be applied to other targeting signals to advance the prediction techniques in that area. Also, subcellular localizations information can be further integrated into

SPdb and to create a “Cell-wide Targeting Signals Map” of the different proteins that have been sequenced and annotated. Such work could help to provide yet another perspective to investigate SPs and to impact on other areas.

The work in *Chapter 5* represents the author’s effort to generate a theoretical model of the SP-SPase-complex. The technique can be repeated to other known SPs to investigate how they come in contact with the receptor. Also, it would be interesting to investigate the conformation upon mutation residues in the structure just as what was being done by using site-directed mutagenesis (Karla *et al.*, 2005). This can gives us a clue of how the different SPs can be bounded. This could contribute to the effort of providing a novel target for antibiotic design. Further, with the availability of more 3D structures, particularly if a SPase-SP-complex is available, it would be interesting to compare those structures with the current model.

In *Chapter 6*, we have developed SNIPn, a new prediction method for the prediction of SP that has achieved good accuracy, particularly in the bacterial datasets. Since the method exploited structural features of SPs, it can be applied to the so called “non-classical SPs”, which do not bear any sequence motifs. Furthermore, although we have conducted preliminary testing on eukaryotic, archaeal and viral (data not shown) sequences, the results are not comparable to the results in bacterial datasets. This is because SNIPn currently uses the Gram- bacteria (*E. coli*) structure template to model for all organisms. With the availability of more sequences and 3D structure data, it will be interesting to apply the same technique to re-model for each organism group and to examine the respective accuracy. The concept of exploiting sequence and structural information could also be extended as a possible means to study other targeting signals since such signal sequences do undergo the recognition and cleavage processes.

For now, SNIPn can be useful in certain domains. One example is to use it for predicting the start of a protein sequence as it is often a challenging task since the beginning of the sequence depends on various elements such as promoters, splicing and so on. Another use is to identify potential (novel) secreted proteins in genomic studies, thus adding yet another tool for the discovery of novel secreted biomarkers (Diamandis, 2004; Xue *et al.*, 2008). Also, many experimental SPs-related studies often alter amino acid at various positions along the SP or portion of the MP through site-directed mutagenesis to study the effect. This is similarly conducted in works that seek to design efficient, or optimize existing SPs or sometimes simply to design synthetic functional SPs. These works are often complex and time-consuming where a multitude of parameters have to be varied while keeping the overall properties in balance (Jain *et al.*, 1994). The prediction method developed in this work can aid in modeling virtual constructs or simply to serve as a preliminary tool for the verification of a potential SP.

7.4 Publications and Presentations Summary

The work described in this thesis has been published in several international peer-reviewed journals and a book chapter as a co-author. Our paper titled "*SPdb: A signal peptide database*" was designated "Highly accessed" by BMC Bioinformatics. Various parts of this work were presented at several national and international conferences or symposiums. The paper titled "*A comprehensive assessment of N-terminal signal peptides prediction methods*" was awarded Best Paper Award at the International Conference on Bioinformatics (InCoB) 2009.

7.4.1 Journal papers

1. Choo, K. H., Tan, T. W., Ranganathan, S., 2009. A comprehensive assessment of N-terminal signal peptides prediction methods. *BMC Bioinformatics.* **10**(15):S2.
2. Choo, K. H., Ranganathan, S., 2008. Flanking signal and mature peptide residues influence signal peptide cleavage. *BMC Bioinformatics*. **9**(12):S15.
3. Choo, K. H., Tong, J. C., Ranganathan, S., 2008. Modeling *Escherichia coli* signal peptidase complex with bound substrate: determinants in the mature peptide influencing signal peptide cleavage. *BMC Bioinformatics*. **9**(1):S15.
4. Choo, K. H., Tan, T. W., Ranganathan, S., 2005. SPdb: A signal peptide database. *BMC Bioinformatics*, **6**:249-257.
5. Choo, K. H., Tong, J. C., Zhang, L. X., 2004. Recent applications of hidden Markov models in computational biology – A Review. *Genomics, Proteomics & Bioinformatics*, 2: 84-96.

7.4.2 Book chapter

1. Tan, T. W., Choo, K. H., Tong, J. C., Tammi, M. T., Bajic, V., 2004. Biological databases and web services: metrics for qualitative analysis. In: *Information Processing and Living Systems*. Edited by Tan, T. W. and Bajic, V. World Scientific Publishing Co., vol. **2**. World Scientific Publishing Co., pp. 771-778. (The findings and insights obtained from this work were incorporated and applied to the development of the pipeline for generating SPdb)

7.4.3 Oral presentations

1. 1st Biochemistry Student Symposium, Sep 2008, Singapore. Signal peptide and its adjacent residues.
2. 1st Symposium on Computational Biology (SYMBIO 2008), Aug 2008, Singapore. Slicing and dicing bacterial and eukaryotic amino-terminal targeting signals through bioinformatics gadgetry.
3. Pre-18th The Federation of Asian and Oceanian Biochemists and Molecular Biologists (FAOBMB) symposium satellite workshop on bioinformatics, Nov 2005, Lahore, Pakistan. Automating biological database creation.
4. 1st Association for Medical and Bio-Informatics, Singapore (AMBIS) bioinformatics symposium, Aug 2003, Singapore. Signal peptide bioinformatics.

7.4.4 Poster presentations

1. The 12th International Conference on Intelligent Systems Molecular Biology (ISMB), Aug 2004, Glasgow, Scotland, UK. SPD: a signal peptide database.
2. The 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB), Mar 2004, San Diego, USA. Protein family classification using Support Vector Machines.

Bibliography

Abagyan, R. and Totrov, M. 1999. *Ab initio* folding of peptides by the optimal-bias Monte Carlo minimization procedure. *J Comput Phys*, **151**(1):402-421.

Abagyan, R., Totrov, M. and Kuznetsov, D. 2004. ICM – a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J Comp Chem*, **15**(5):488-506.

Adams, H., Scotti, P. A., De Cock, H., Luirink, J. and Tommassen, J. 2002. The presence of a helix breaker in the hydrophobic core of signal sequences of secretory proteins prevents recognition by the signal-recognition particle in *Escherichia coli*. *Eur J Biochem*, **269**(22):5564-5571

Ahn, J. H., Hwang, M. Y., Lee, K. H., Choi, C. Y. and Kim, D. M. 2007. Use of signal sequences as an in situ removable sequence element to stimulate protein synthesis in cell-free extracts. *Nucleic Acids Res*, **35**(4):e21.

Akutsu, T. and Sim, K. L. 1999. Protein threading based on multiple protein structure alignment. *Genome Inform Ser Workshop Genome Inform*, **10**:23-29.

Al-Qahtani, A., Teilhet, M. and Mensa-Wilmot, K. 1998. Species-specificity in endoplasmic reticulum signal peptide utilization revealed by proteins from *Trypanosoma brucei* and *Leishmania*. *Biochem J*, **331 (Pt 2)**:521-529.

Albers, S. V., Konings, W. N. and Driessen, A. J. 1999. A unique short signal sequence in membrane-anchored proteins of Archaea. *Mol Microbiol*, **31**(5):1595-1596.

Alken, M., Rutz, C., Kochl, R., Donalies, U., Oueslati, M., Furkert, J., Wietfeld, D., Hermosilla, R., Scholz, A., Beyermann, M., Rosenthal, W. and Schulein, R. 2005. The signal peptide of the rat corticotropin-releasing factor receptor 1 promotes receptor expression but is not essential for establishing a functional receptor. *Biochem J*, **390**(Pt 2):455-464.

Allet, B., Bernard, A. R., Hochmann, A., Rohrbach, E., Graber, P., Magnenat, E., Mazzei, G. J. and Bernasconi, L. 1997. A bacterial signal peptide directs efficient secretion of eukaryotic proteins in the baculovirus expression system. *Protein Expr Purif*, **9**(1):61-68.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17):3389-3402.

Andersson, H. and von Heijne, G. 1991. A 30-residue-long "export initiation domain" adjacent to the signal sequence is critical for protein translocation across the inner membrane of *Escherichia coli*. *Proc Natl Acad Sci U S A*, **88**(21):9751-9754.

Andrews, D. W., Perara, E., Lesser, C. and Lingappa, V. R. 1988. Sequences beyond the cleavage site influence signal peptide function. *J Biol Chem*, **263**(30):15791-15798.

Anjos, S., Nguyen, A., Ounissi-Benkalha, H., Tessier, M. C. and Polychronakos, C. 2002. A common autoimmunity predisposing signal peptide variant of the cytotoxic T-lymphocyte antigen 4 results in inefficient glycosylation of the susceptibility allele. *J Biol Chem*, **277**(48):46478-46486.

Antelmann, H., Tjalsma, H., Voigt, B., Ohlmeier, S., Bron, S., van Dijl, J. M. and Hecker, M. 2001. A proteomic view on genome-based signal peptide predictions. *Genome Res*, **11**(9):1484-1502.

Applied Biosystems. 2008. http://www.labtechnologist.com/Products/Applied-Bio-sequences-a-human-genome-for-60-000.

Asayama, S., Kawamura, E., Nagaoka, S. and Kawakami, H. 2006. Design of manganese porphyrin modified with mitochondrial signal peptide for a new antioxidant. *Mol Pharm*, **3**(4):468-470.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**(1):25-29.

Austen, B. M., Hermon-Taylor, J., Kaderbhai, M. A. and Ridd, D. H. 1984. Design and synthesis of a consensus signal sequence that inhibits protein translocation into rough microsomal vesicles. *Biochem J*, **224**(1):317-325.

Baier, M., Bannert, N., Werner, A., Lang, K. and Kurth, R. 1997. Molecular cloning, sequence, expression, and processing of the interleukin 16 precursor. *Proc Natl Acad Sci U S A*, **94**(10):5273-5277.

Bagos, P. G., Tsirigos, K. D., Plessas, S. K., Liakopoulos, T. D. and Hamodrakas, S. J. 2009. Prediction of signal peptides in archaea. *Protein Engineering, Design and Selection*, **22**(1):27-35.

Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E. 2004. Swiss-Prot: juggling between evolution and stability. *Brief Bioinform*, **5**(1):39-55.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. and Yeh, L. S. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res*, **33**(Database issue):D154-159.

Bankaitis, V. A. and Bassford, P. J., Jr. 1985. Sequences within the mature maltose-binding protein of *Escherichia coli* may be actively involved in initiating the export process. *Ann Inst Pasteur Microbiol*, **136B**(1):3-7.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. and Miyano, S. 2002. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**(2):298-305.

Barash, S., Wang, W. and Shi, Y. 2002. Human secretory signal peptide description by hidden Markov model and generation of a strong artificial signal peptide for secreted protein expression. *Biochem Biophys Res Commun*, **294**(4):835-842.

Bardy, S. L., Eichler, J. and Jarrell, K. F. 2003. Archaeal signal peptides--a comparative survey at the genome level. *Protein Sci*, **12**(9):1833-1843.

Bardy, S. L., Ng, S. Y., Carnegie, D. S. and Jarrell, K. F. 2005. Site-directed mutagenesis analysis of amino acids critical for activity of the type I signal peptidase of the archaeon *Methanococcus voltae*. *J Bacteriol*, **187**(3):1188-1191.

Bagos, P. G., Tsirigos, K. D., Plessas, S. K., Liakopoulos, T. D. and Hamodrakas, S. J. 2008. Prediction of signal peptides in archaea. *Protein Engineering, Design and Selection*, **22**(1):27-35.

Barkocy-Gallagher, G. A. and Bassford, P. J., Jr. 1992. Synthesis of precursor maltose-binding protein with proline in the +1 position of the cleavage site interferes with the activity of *Escherichia coli* signal peptidase I in vivo. *J Biol Chem*, **267**(2):1231-1238.

Batey, R. T., Rambo, R. P., Lucast, L., Rha, B. and Doudna, J. A. 2000. Crystal structure of the ribonucleoprotein core of the signal recognition particle. *Science*, **287**(5456):1232-1239.

Beckmann, R., Bubeck, D., Grassucci, R., Penczek, P., Verschoor, A., Blobel, G. and Frank, J. 1997. Alignment of conduits for the nascent polypeptide chain in the ribosome-Sec61 complex. *Science*, **278**(5346):2123-2126.

Belin, D., Guzman, L. M., Bost, S., Konakova, M., Silva, F. and Beckwith, J. 2004. Functional activity of eukaryotic signal sequences in *Escherichia coli*: the ovalbumin family of serine protease inhibitors. *J Mol Biol*, **335**(2):437-453.

Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G. and Brunak, S. 2004a. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel*, **17**(4):349-356.

Bendtsen, J. D., Nielsen, H., von Heijne, G. and Brunak, S. 2004b. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**(4):783-795.

Bendtsen, J. D., Binnewies, T. T., Hallin, P. F., Sicheritz-Ponten, T. and Ussery, D. W. 2005a. Genome update: prediction of secreted proteins in 225 bacterial proteomes. *Microbiology*, **151**(Pt 6):1725-1727.

Bendtsen, J. D., Kiemer, L., Fausboll, A. and Brunak, S. 2005b. Non-classical protein secretion in bacteria. *BMC Microbiol*, **5**:58.

Berks, B. C. 1996. A common export pathway for proteins binding complex redox cofactors? *Mol Microbiol*, **22**(3):393-404.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res*, **28**(1):235-242.

Bernstein, H. D. 1998. Protein targeting: getting into the groove. *Curr Biol*, **8**(20):R715-718.

Bird, P., Gething, M. J. and Sambrook, J. 1987. Translocation in yeast and mammalian cells: not all signal sequences are functionally equivalent. *J Cell Biol*, **105**(6):2905-2914

Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyras, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H. R., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodwark, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T. and Clamp, M. 2004. An overview of Ensembl. *Genome Res*, **14**(5):925-928.

Biro, J. C. 2006. Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theor Biol Med Model*, **3**:15.

Bjellqvist, B., Hughes, G. J., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J. C., Frutiger, S. and Hochstrasser, D. 1993. The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis*, **14**(10):1023-1031.

Black, M. T. and Bruton, G. 1998. Inhibitors of bacterial signal peptidases. *Curr Pharm Des*, **4**(2):133-154.

Blanco, D. R., Whitelegge, J. P., Miller, J. N. and Lovett, M. A. 1999. Demonstration by mass spectrometry that purified native *Treponema pallidum* rare outer membrane protein 1 (Tromp1) has a cleaved signal peptide. *J Bacteriol*, **181**(16):5094-5098.

Bland, F. A., Lemberg, M. K., McMichael, A. J., Martoglio, B. and Braud, V. M. 2003. Requirement of the proteasome for the trimming of signal peptide-derived epitopes presented by the nonclassical major histocompatibility complex class I molecule HLA-E. *J Biol Chem*, **278**(36):33747-33752.

Blaudeck, N., Sprenger, G. A., Freudl, R. and Wiegert, T. 2001. Specificity of signal peptide recognition in tat-dependent bacterial protein translocation. *J Bacteriol*, **183**(2):604-610.

Blobel, G. and Dobberstein, B. 1975a. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *J Cell Biol*, **67**(3):835-851.

Blobel, G. and Dobberstein, B. 1975b. Transfer of proteins across membranes. II. Reconstitution of functional rough microsomes from heterologous components. *J Cell Biol*, **67**(3):852-862.

Blobel, G. and Sabatini, D. D. 1971. Ribosome-membrane interaction in eukaryotic cells. *Biomembranes*, **2**:193-195.

Blobel, G. 2000. Nobel Lecture on Protein Targeting. *Bioscience Reports*, 20(5):303-344.

Bodén, M. and Hawkins, J. 2005. Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics*, **21**(10):2279-2286.

Bogsch, E., Brink, S. and Robinson, C. 1997. Pathway specificity for a delta *pH*-dependent precursor thylakoid lumen protein is governed by a 'Sec-avoidance' motif in the transfer peptide and a '*Sec*-incompatible' mature protein. *Embo J*, **16**(13):3851-3859.

Bonfanti, R., Colombo, C., Nocerino, V., Massa, O., Lampasona, V., Iafusco, D., Viscardi, M., Chiumello, G., Meschi, F. and Barbetti, F. 2009. Insulin gene mutations as cause of diabetes in children negative for five type 1 diabetes autoantibodies. *Diabetes Care*, **32**(1):123-125.

Bonin-Debs, A. L., Boche, I., Gille, H. and Brinkmann, U. 2004. Development of secreted proteins as biotherapeutic agents. *Expert Opin Biol Ther*, **4**(4):551-558.

Bork, P. 2000. Powers and pitfalls in sequence analysis: the 70% hurdle. *Genome Res*, **10**(4):398-400.

Bowden, G. A., Baneyx, F. and Georgiou, G. 1992. Abnormal fractionation of beta-lactamase in *Escherichia coli*: evidence for an interaction with the inner membrane in the absence of a leader peptide. *J Bacteriol*, **174**(10):3407-3410.

Bradford, J. R. 2001. In: MRes Bioinformatics thesis "*In silico* Methods for Prediction of Signal Peptides and their Cleavage Sites, and Linear Epitopes". School of Biochemistry and Molecular Biology, The University of Leeds.

Bradshaw, N. and Walter, P. 2007. The signal recognition particle (SRP) RNA links conformational changes in the SRP to protein targeting. *Mol Biol Cell*, **18**(7):2728-2734.

Bradshaw, N., Neher, S. B., Booth, D. S. and Walter, P. 2009. Signal sequences activate the catalytic switch of SRP RNA. *Science*, **323**(5910):127-130.

Braud, V., Jones, E. Y. and McMichael, A. 1997. The human major histocompatibility complex class Ib molecule HLA-E binds signal sequence-derived peptides with primary anchor residues at positions 2 and 9. *Eur J Immunol*, **27**(5):1164-1169.

Braud, V. M., Allan, D. S., O'Callaghan, C. A., Soderstrom, K., D'Andrea, A., Ogg, G. S., Lazetic, S., Young, N. T., Bell, J. I., Phillips, J. H., Lanier, L. L. and McMichael, A. J. 1998. HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature*, **391**(6669):795-799.

Briggs, M. S., Gierasch, L. M., Zlotnick, A., Lear, J. D. and DeGrado, W. F. 1985. In vivo function and membrane binding properties are correlated for *Escherichia coli lamB* signal peptides. *Science*, **228**(4703):1096-1099.

Brockmeier, U., Caspers, M., Freudl, R., Jockwer, A., Noll, T. and Eggert, T. 2006. Systematic screening of all signal peptides from *Bacillus subtilis:* a powerful strategy in optimizing heterologous protein secretion in Gram-positive bacteria. *J Mol Biol*, **362**(3):393-402.

Brusic, V. 2007. The growth of bioinformatics. *Brief Bioinform*, **8**(2):69-70.

Burges, C. 1998. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc*, **2**(2):121-167.

Burghaus, P. A. and Lingelbach, K. 2001. Luciferase, when fused to an N-terminal signal peptide, is secreted from transfected *Plasmodium falciparum* and transported to the cytosol of infected erythrocytes. *J Biol Chem*, **276**(29):26838-26845.

Buus, S. 1999. Description and prediction of peptide-MHC binding: the human MHC project. *Curr Opin Immunol*, **11**(2):209-213.

Buzder-Lantos, P., Bockstael, K., Anné, J., Herdewijn, P. 2009. Substrate based peptide aldehyde inhibits bacterial type I signal peptidase. *Bioorg Med Chem Let*. (*in press*).

Cai, Y. D., Lin, S. L. And Chou, K. C. 2003. Support vector machines for prediction of protein signal sequences and their cleavage sites. *Peptides*, **24**:159-161.

Carlos, J. L., Paetzel, M., Brubaker, G., Karla, A., Ashwell, C. M., Lively, M. O., Cao, G., Bullinger, P. and Dalbey, R. E. 2000. The role of the membrane-spanning domain of type I signal peptidases in substrate cleavage site selection. *J Biol Chem*, **275**(49):38813-38822.

Chan, D., Ho, M. S. and Cheah, K. S. 2001. Aberrant signal peptide cleavage of collagen X in Schmid metaphyseal chondrodysplasia. Implications for the molecular basis of the disease. *J Biol Chem*, **276**(11):7992-7997.

Chang, C. C. and Lin, C. J. 2001. LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, L., Tai, P. C., Briggs, M. S. and Gierasch, L. M. 1987. Protein translocation into *Escherichia coli* membrane vesicles is inhibited by functional synthetic signal peptides. *J Biol Chem*, **262**(4):1427-1429.

Chen, H. and Leder, P. 1999. A new signal sequence trap using alkaline phosphatase as a reporter. *Nucleic Acids Res*, **27**(4):1219-1222.

Chen, Y., Yu, P., Luo, J. and Jiang, Y. 2003. Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. *Mamm Genome*, **14**(12):859-865.

Chen, Y., Zhang, Y., Yin, Y., Gao, G., Li, S., Jiang, Y., Gu, X. and Luo, J. 2005. SPD--a web-based secreted protein database. *Nucleic Acids Res*, **33**(Database issue):D169-173.

Choo, K. H., Tong, J. C. and Zhang, L. 2004. Recent applications of Hidden Markov Models in computational biology. *Genomics Proteomics Bioinformatics*, **2**(2):84-96.

Choo, K. H., Tan, T. W. and Ranganathan, S. 2005. SPdb--a signal peptide database. *BMC Bioinformatics*, **6**:249.

Choo, K. H. and Ranganathan, S. 2008. Flanking signal and mature peptide residues influence signal peptide cleavage. *BMC Bioinformatics*, **9**(12):S15.

Chou, K. C. 2001a. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*, **43**(3):246-255.

Chou, K. C. 2001b. Prediction of protein signal sequences and their cleavage sites. *Proteins*, **42**(1):136-139.

Chou, K. C. and Shen, H. B. 2007. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Commun*, **357**(3):633-640.

Clark, H. F., Gurney, A. L., Abaya, E., Baker, K., Baldwin, D., Brush, J., Chen, J., Chow, B., Chui, C., Crowley, C., Currell, B., Deuel, B., Dowd, P., Eaton, D., Foster, J., Grimaldi, C., Gu, Q., Hass, P. E., Heldens, S., Huang, A., Kim, H. S., Klimowski, L., Jin, Y., Johnson, S., Lee, J., Lewis, L., Liao, D., Mark, M., Robbie, E., Sanchez, C., Schoenfeld, J., Seshagiri, S., Simmons, L., Singh, J., Smith, V., Stinson, J., Vagts, A., Vandlen, R., Watanabe, C., Wieand, D., Woods, K., Xie, M. H., Yansura, D., Yi, S., Yu, G., Yuan, J., Zhang, M., Zhang, Z., Goddard, A., Wood, W. I., Godowski, P. and Gray, A. 2003. The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. *Genome Res*, **13**(10):2265-2270.

Clemons, W. M., Jr., Gowda, K., Black, S. D., Zwieb, C. and Ramakrishnan, V. 1999. Crystal structure of the conserved subdomain of human protein SRP54M at 2.1 A resolution: evidence for the mechanism of signal peptide binding. *J Mol Biol*, **292**(3):697-705.

Cleverley, R. M. and Gierasch, L. M. 2002. Mapping the signal sequence-binding site on SRP reveals a significant role for the NG domain. *J Biol Chem*, **277**(48):46763-46768.Bird, P., Gething, M. J. and Sambrook, J. 1987. Translocation in yeast and mammalian cells: not all signal sequences are functionally equivalent. *J Cell Biol*, **105**(6 Pt 2):2905-2914.

Corney, D. P., Buxton, B. F., Langdon, W. B. and Jones, D. T. 2004. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, **20**(17):3206-3213.

Corradin, G. and Demotz, S. 1997. Peptide-MHC complexes assembled following multiple pathways: an opportunity for the design of vaccines and therapeutic molecules. *Hum Immunol*, **54**(2):137-147.

Crawshaw, S. G., Martoglio, B., Meacock, S. L. and High, S. 2004. A misassembled transmembrane domain of a polytopic protein associates with signal peptide peptidase. *Biochem J*, **384**(Pt 1):9-17.

Creuzenet, C., Durand, C. and Haertle, T. 1997. Interaction of alpha s2- and beta-casein signal peptides with DMPC and DMPG liposomes. *Peptides*, **18**(4):463-472.

Crooks, G. E., Hon, G., Chandonia, J. M. and Brenner, S. E. 2004. WebLogo: a sequence logo generator. *Genome Res*, **14**(6):1188-1190.

Crowley, K. S., Liao, S., Worrell, V. E., Reinhart, G. D. and Johnson, A. E. 1994. Secretory proteins move through the endoplasmic reticulum membrane via an aqueous, gated pore. *Cell*, **78**(3):461-471.

Dalbey, R. E. and Von Heijne, G. 1992. Signal peptidases in prokaryotes and eukaryotes--a new protease family. *Trends Biochem Sci*, **17**(11):474-478.

Dalbey, R. E., Kuhn, A. and von Heijne, G. 1995. Directionality in protein translocation across membranes: the N-tail phenomenon. *Trends Cell Biol*, **5**(10):380-383.

Dalbey, R. E., Lively, M. O., Bron, S. and van Dijl, J. M. 1997. The chemistry and enzymology of the type I signal peptidases. *Protein Sci*, **6**(6):1129-1138.

Dalbey, R. E. and von Heijne, G. 2002. Protein targeting, transport & translocation. Academic Press, 1st edn. ISBN-10: 012200731X.

Date, T. 1983. Demonstration by a novel genetic technique that leader peptidase is an essential enzyme of *Escherichia coli*. *J Bacteriol*, **154**(1):76-83.

Datta, R., Waheed, A., Shah, G. N. and Sly, W. S. 2007. Signal sequence mutation in autosomal dominant form of hypoparathyroidism induces apoptosis that is corrected by a chemical chaperone. *Proc Natl Acad Sci U S A*, **104**(50):19989-19994.

de Castro, E., Sigrist, C. J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A. and Hulo, N. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*, **34**(Web Server issue):W362-365.

de Gier, J. W., Scotti, P. A., Saaf, A., Valent, Q. A., Kuhn, A., Luirink, J. and von Heijne, G. 1998. Differential use of the signal recognition particle translocase targeting pathway for inner membrane protein assembly in Escherichia coli. *Proc Natl Acad Sci U S A*, **95**(25):14646-14651.

Della-Cioppa, G., Kishore, G. M., Beachy, R. N. and Fraley, R. T. 1987. Protein Trafficking in Plant Cells. *Plant Physiol*, **84**(4):965-968.

Deshaies, R. J. and Schekman, R. 1989. SEC62 encodes a putative membrane protein required for protein translocation into the yeast endoplasmic reticulum. *J Cell Biol*, **109**(6 Pt 1):2653-2664.

Diamandis, E. P. 2004. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics*, **3**(4):367-378.

Ding, B., Kull, B., Liu, Z., Mottagui-Tabar, S., Thonberg, H., Gu, H. F., Brookes, A. J., Grundemar, L., Karlsson, C., Hamsten, A., Arner, P., Ostenson, C. G., Efendic, S., Monne, M., von Heijne, G., Eriksson, P. and Wahlestedt, C. 2005. Human neuropeptide Y signal peptide gain-of-function polymorphism is associated with increased body mass index: possible mode of function. *Regul Pept*, **127**(1-3):45-53.

Driessen, A. J. and van der Does, C. 2002. Protein export in bacteria. In: *Protein targeting, transport & translocation.* Edited by Dalbey, R. E. and von Heijne, G. Academic Press, 1st edn, pp. 47-73. ISBN-10: 012200731X.

Duffaud, G. and Inouye, M. 1988. Signal peptidases recognize a structural feature at the cleavage site of secretory proteins. *J Biol Chem*, **263**(21):10224-10228.

Dultz, E., Hildenbeutel, M., Martoglio, B., Hochman, J., Dobberstein, B. and Kapp, K. 2008. The signal peptide of the mouse mammary tumor virus *Rem* protein is released from the endoplasmic reticulum membrane and accumulates in nucleoli. *J Biol Chem*, **283**(15):9966-9976.

Ebel, T., Gerhards, J., Binder, B. R. and Lipp, J. 1999. *Theileria parva* 104 kDa microneme--rhoptry protein is membrane-anchored by a non-cleaved amino-terminal signal sequence for entry into the endoplasmic reticulum. *Mol Biochem Parasitol*, **100**(1):19-26.

Ebel, T., Pelle, R., Janoo, R., Lipp, J. and Bishop, R. 2004. A membrane-anchored *Theileria parva* cyclophilin with a non-cleaved amino-terminal signal peptide for entry into the endoplasmic reticulum. *Vet Parasitol*, **121**(1-2):65-77.

Eddy, S. R. 1998. Profile hidden Markov models. *Bioinformatics*, **14**(9):755-763.

Eichler, R., Lenz, O., Strecker, T. and Garten, W. 2003. Signal peptide of Lassa virus glycoprotein GP-C exhibits an unusual length. *FEBS Lett*, **538**(1-3):203-206.

Eidhammer, I., Jonassen, I. and Taylor, W. R. 2004. Protein bioinformatics: an algorithmic approach to sequence and structure analysis. Wiley; 1 edn. ISBN-10: 0470848391.

Eisenberg, D., Weiss, R. M. and Terwilliger, T. C. 1982. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, **299**(5881):371-374.

Ekici, O. D., Karla, A., Paetzel, M., Lively, M. O., Pei, D. and Dalbey, R. E. 2007. Altered -3 substrate specificity of *Escherichia coli* signal peptidase 1 mutants as revealed by screening a combinatorial peptide library. *J Biol Chem*, **282**(1):417-425.

Elling, A. A., Mitreva, M., Gai, X., Martin, J., Recknor, J., Davis, E. L., Hussey, R. S., Nettleton, D., McCarter, J. P., Baum, T. J. 2009. Sequence mining and transcript profiling to explore cyst nematode parasitism. *BMC Genomics*, **10**:58.

Emanuelsson, O., Nielsen, H. and von Heijne, G. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci*, **8**(5):978-984.

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol*, **300**(4):1005-1016.

Eusebio, A., Friedberg, T. and Spiess, M. 1998. The role of the hydrophobic domain in orienting natural signal sequences within the ER membrane. *Exp Cell Res*, **241**(1):181-185.

Fariselli, P., Finocchiaro, G. and Casadio, R. 2003. SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. *Bioinformatics*, **19**(18):2498-2499.

Fekkes, P. and Driessen, A. J. 1999. Protein targeting to the bacterial cytoplasmic membrane. *Microbiol Mol Biol Rev*, **63**(1):161-173.

Feldheim, D. and Schekman, R. 1994. Sec72p contributes to the selective recognition of signal peptides by the secretory polypeptide translocation complex. *J Cell Biol*, **126**(4):935-943.

Fernandez-Recio, J., Totrov, M. and Abagyan, R. 2002. Soft protein-protein docking in internal coordinates. *Protein Sci*, **11**(2):280-291.

Fikes, J. D., Barkocy-Gallagher, G. A., Klapper, D. G. and Bassford, P. J., Jr. 1990. Maturation of Escherichia coli maltose-binding protein by signal peptidase I *in vivo*. Sequence requirements for efficient processing and demonstration of an alternate cleavage site. *J Biol Chem*, **265**(6):3417-3423.

Fingerhut, A., Reutrakul, S., Knuedeler, S.D., Moeller, L.C., Greenlee, C., Refetoff, S. and Janssen, O.E. (2004) Partial deficiency of thyroxine-binding globulin-allentown is due to a mutation in the signal peptide. *J Clin Endocrinol Metab*, **89**, 2477–2483.

Flower, A. M., Doebele, R. C. and Silhavy, T. J. 1994. *PrlA* and *PrlG* suppressors reduce the requirement for signal sequence recognition. *J Bacteriol*, **176**(18):5607-5614.

Folz, R. J., Nothwehr, S. F. and Gordon, J. I. 1988. Substrate specificity of eukaryotic signal peptidase. Site-saturation mutagenesis at position -1 regulates cleavage between multiple sites in human pre (delta pro) apolipoprotein A-II. *J Biol Chem*, **263**(4):2070-2078.

Frank, K. and Sippl, M. J. 2008. High-performance signal peptide prediction based on sequence alignment techniques. *Bioinformatics*, **24**(19):2172-2176.

Frate, M. C., Lietz, E. J., Santos, J., Rossi, J. P., Fink, A. L. and Ermacora, M. R. 2000. Export and folding of signal-sequenceless *Bacillus licheniformis* beta-lactamase in *Escherichia coli*. *Eur J Biochem*, **267**(12):3836-3847.

Froeschke, M., Basler, M., Groettrup, M. and Dobberstein, B. 2003. Long-lived signal peptide of lymphocytic choriomeningitis virus glycoprotein pGP-C. *J Biol Chem*, **278**(43):41914-41920.

Gaizauskas, R., Demetriou, G., Artymiuk, P. J. and Willett, P. 2003. Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, **19**(1):135-143.

Gavel, Y. and von Heijne, G. 1990. A conserved cleavage-site motif in chloroplast transit peptides. *FEBS Lett*, **261**(2):455-458.

Gennity, J., Goldstein, J. and Inouye, M. 1990. Signal peptide mutants of *Escherichia coli*. *J Bioenerg Biomembr*, **22**(3):233-269.

Georgiou, G. and Segatori, L. 2005. Preparative expression of secreted proteins in bacteria: status report and future prospects. *Curr Opin Biotechnol*, **16**(5):538-545.

Gierasch, L. M. 1989. Signal sequences. *Biochemistry*, **28**(3):923-930.

Gill, D. R. and Salmond, G. P. 1990. The identification of the *Escherichia coli ftsY* gene product: an unusual protein. *Mol Microbiol*, **4**(4):575-583.

Gilmore, R., Blobel, G. and Walter, P. 1982a. Protein translocation across the endoplasmic reticulum. I. Detection in the microsomal membrane of a receptor for the signal recognition particle. *J Cell Biol*, **95**(2 Pt 1):463-469.

Gilmore, R., Walter, P. and Blobel, G. 1982b. Protein translocation across the endoplasmic reticulum. II. Isolation and characterization of the signal recognition particle receptor. *J Cell Biol*, **95**(2 Pt 1):470-477.

Goder, V. and Spiess, M. 2003. Molecular mechanism of signal sequence orientation in the endoplasmic reticulum. *Embo J*, **22**(14):3645-3653.

Gomi, M., Sonoyama, M. and Mitaku, S. 2004. High performance system for signal peptide prediction: SOSUIsignal. *Chem-Bio Info J*, **4**:142-147.

Grabley, S. and Thiericke, R. 1999. Bioactive agents from natural sources: trends in discovery and application. *Adv Biochem Eng Biotechnol*, **64**:101-154.

Guo, X., Zhang, Y., Zhang, X., Wang, S. and Lu, C. 2008. Recognition of signal peptide by protein translocation machinery in middle silk gland of silkworm *Bombyx mori*. *Acta Biochim Biophys Sin (Shanghai)*, **40**(1):38-46.

Haigh, N. G. and Johnson, A. E. 2002. A new role for BiP: closing the aqueous translocon pore during protein integration into the ER membrane. *J Cell Biol*, **156**(2):261-270.

Hamman, B. D., Chen, J. C., Johnson, E. E. and Johnson, A. E. 1997. The aqueous pore through the translocon has a diameter of 40-60 A during cotranslational protein translocation at the ER membrane. *Cell*, **89**(4):535-544.

Hanein, D., Matlack, K. E., Jungnickel, B., Plath, K., Kalies, K. U., Miller, K. R., Rapoport, T. A. and Akey, C. W. 1996. Oligomeric rings of the Sec61p complex induced by ligands required for protein translocation. *Cell*, **87**(4):721-732.

Hardy, S. J. and Randall, L. L. 1991. A kinetic partitioning model of selective binding of nonnative proteins by the bacterial chaperone *SecB*. *Science*, **251**(4992):439-443.

Harwood, C. R. and Cranenburgh, R. 2008. *Bacillus* protein secretion: an unfolding story. *Trends Microbiol*, **16**(2):73-79.

Hawkins, J. and Bodén, M. 2006. Detecting and sorting targeting peptides with neural networks and support vector machines. *J Bioinform Comput Biol*, **4**(1):1-18.

Hegde, R. S. and Bernstein, H. D. 2006. The surprising complexity of signal sequences. *Trends Biochem Sci*, **31**(10):563-571.

Henderson, R. A., Michel, H., Sakaguchi, K., Shabanowitz, J., Appella, E., Hunt, D. F. and Engelhard, V. H. 1992. HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. *Science*, **255**(5049):1264-1266.

Henikoff, J. G. and Henikoff, S. 1996. Using substitution probabilities to improve position-specific scoring matrices. *Comput Appl Biosci*, **12**(2):135-143.

Hiller, K., Grote, A., Scheer, M., Munch, R. and Jahn, D. 2004. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res*, **32**(Web Server issue):W375-379.

Himeno, T., Imanaka, T. and Aiba, S. 1986. Protein secretion in Bacillus subtilis as influenced by the combination of signal sequence and the following mature portion. *FEMS Microbiol Lett*, **35**(1):17-21.

Hinuma, S., Habata, Y., Fujii, R., Kawamata, Y., Hosoya, M., Fukusumi, S., Kitada, C., Masuo, Y., Asano, T., Matsumoto, H., Sekiguchi, M., Kurokawa, T., Nishimura, O., Onda, H. and Fujino, M. 1998. A prolactin-releasing peptide in the brain. *Nature*, **393**(6682):272-276.

Holland, I. B. 2004. Translocation of bacterial proteins--an overview. *Biochim Biophys Acta*, **1694**(1-3):5-16.

Hombach, J., Pircher, H., Tonegawa, S. and Zinkernagel, R. M. 1995. Strictly transporter of antigen presentation (TAP)-dependent presentation of an immunodominant cytotoxic T lymphocyte epitope in the signal sequence of a virus protein. *J Exp Med*, **182**(5):1615-1619.

Hon, L. S., Zhang, Y., Kaminker, J. S. and Zhang, Z. 2009. Computational prediction of the functional effects of amino acid substitutions in signal peptides using a model-based approach. *Hum Mutat*, **30**(1):99-106.

Hope, R. G., McElwee, M. J. and McLauchlan, J. 2006. Efficient cleavage by signal peptide peptidase requires residues within the signal peptide between the core and E1 proteins of hepatitis C virus strain J1. *J Gen Virol*, **87**(Pt 3):623-627.

Hsu, C. W., Chang, C. C. and Lin, C. J. 2008. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

Hunter, E. 2007. Virus assembly: nuclear localization signals. In: *Fields Virology.* Edited by Knipe, D. M. and Howley, P. M., vol. 1, 5th edn. Lippincott Williams & Wilkins, pp. 143-144.

Ikai, A. 1980. Thermostability and aliphatic index of globular proteins. *J Biochem*, **88**(6):1895-1898.

Ito, M., Oiso, Y., Muraw, T., Kondo, K., Saito, H., Chinzei, T., Racchi, M. and Lively, M.O. (1993) Possible involvement of inefficient cleavage of preprovasopressin by signal peptidase as a cause for familial central diabetes insipidus. *J Clin Invest*, **91**, 2565–2571.

Izard, J. W. and Kendall, D. A. 1994. Signal peptides: exquisitely designed transport promoters. *Mol Microbiol*, **13**(5):765-773.

Izard, J., Parker, M. W., Chartier, M., Duche, D. and Baty, D. 1994. A single amino acid substitution can restore the solubility of aggregated colicin A mutants in *Escherichia coli*. *Protein Eng*, **7**(12):1495-1500.

Jacobs, K. A., Collins-Racie, L. A., Colbert, M., Duckett, M., Golden-Fleet, M., Kelleher, K., Kriz, R., LaVallie, E. R., Merberg, D., Spaulding, V., Stover, J., Williamson, M. J. and McCoy, J. M. 1997. A genetic selection for isolating cDNAs encoding secreted proteins. *Gene*, **198**(1-2):289-296.

Jagla, B. and Schuchhardt, J. 2000. Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. *Bioinformatics*, **16**(3):245-250.

Jain, R. G., Rusch, S. L. and Kendall, D. A. 1994. Signal peptide cleavage regions. Functional limits on length and topological implications. *J Biol Chem*, **269**(23):16305-16310.

Jarjanazi, H., Savas, S., Pabalan, N., Dennis, J. W. and Ozcelik, H. 2008. Biological implications of SNPs in signal peptide domains of human proteins. *Proteins*, **70**(2):394-403.

Jiang, C., Magee, D. M., Ivey, F. D. and Cox, R. A. 2002. Role of signal sequence in vaccine-induced protection against experimental coccidioidomycosis. *Infect Immun*, **70**(7):3539-3545.

Joachims, T. 2002. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Springer, 1st edn. ISBN-10: 079237679X

Joliot, A., Maizel, A., Rosenberg, D., Trembleau, A., Dupas, S., Volovitch, M. and Prochiantz, A. 1998. Identification of a signal sequence necessary for the unconventional secretion of *Engrailed* homeoprotein. *Curr Biol*, **8**(15):856-863.

Juncker, A. S., Willenbrock, H., Von Heijne, G., Brunak, S., Nielsen, H. and Krogh, A. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*, **12**(8):1652-1662.

Jungnickel, B. and Rapoport, T. A. 1995. A posttargeting signal sequence recognition event in the endoplasmic reticulum membrane. *Cell*, **82**(2):261-270.

Kaiser, C. A., Preuss, D., Grisafi, P. and Botstein, D. 1987. Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science*, **235**(4786):312-317.

Kajava, A. V., Zolov, S. N., Kalinin, A. E. and Nesmeyanova, M. A. 2000. The net charge of the first 18 residues of the mature sequence affects protein translocation across the cytoplasmic membrane of gram-negative bacteria. *J Bacteriol*, **182**(8):2163-2169.

Kajava, A. V., Zolov, S. N., Pyatkov, K. I., Kalinin, A. E. and Nesmeyanova, M. A. 2002. Processing of *Escherichia coli* alkaline phosphatase. Sequence requirements and possible conformations of the -6 to -4 region of the signal peptide. *J Biol Chem*, **277**(52):50396-50402.

Kalies, K. U. and Hartmann, E. 1998. Protein translocation into the endoplasmic reticulum (ER)--two similar routes with different modes. *Eur J Biochem*, **254**(1):1-5.

Käll, L., Krogh, A. and Sonnhammer, E. L. 2004. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, **338**(5):1027-1036.

Kanagasabai, R, Choo, K. H., Ranganathan, S., Baker, C. J., 2007. A workflow for mutation extraction and structure annotation. *J Bioinform Comput Biol*, (Special Issue: Making Sense of Mutations Requires Knowledge Management), **5**(6):1319-1337.

Kang, S. W., Rane, N. S., Kim, S. J., Garrison, J. L., Taunton, J. and Hegde, R. S. 2006. Substrate-specific translocational attenuation during ER stress defines a pre-emptive quality control pathway. *Cell*, **127**(5):999-1013.

Kantardjieff, K. A. and Rupp, B. 2004. Protein isoelectric point as a predictor for increased crystallization screening efficiency. *Bioinformatics*, **20**(14):2162-2168.

Karamyshev, A. L., Karamysheva, Z. N., Kajava, A. V., Ksenzenko, V. N. and Nesmeyanova, M. A. 1998. Processing of *Escherichia coli* alkaline phosphatase: role of the primary structure of the signal peptide cleavage region. *J Mol Biol*, **277**(4):859-870.

Karla, A., Lively, M. O., Paetzel, M. and Dalbey, R. 2005. The identification of residues that control signal peptidase cleavage fidelity and substrate specificity. *J Biol Chem*, **280**(8):6731-6741.

Karlström, A., Jacobsson, K., Flock, M., Flock, J. I. and Guss, B. 2004. Identification of a novel collagen-like protein, SclC, in *Streptococcus equi* using signal sequence phage display. *Vet Microbiol*, **104**(3-4):179-188.

Kaye, J. 2008. The regulation of direct-to-consumer genetic tests. *Hum Mol Genet*, **17**(R2):R180-183.

Keenan, R. J., Freymann, D. M., Walter, P. and Stroud, R. M. 1998. Crystal structure of the signal sequence binding subunit of the signal recognition particle. *Cell*, **94**(2):181-191.

Kim, S. J., Mitra, D., Salerno, J. R. and Hegde, R. S. 2002. Signal sequences control gating of the protein translocation channel in a substrate-specific manner. *Dev Cell*, **2**(2):207-217.

Kiraly, O., Boulling, A., Witt, H., Le Marechal, C., Chen, J. M., Rosendahl, J., Battaggia, C., Wartmann, T., Sahin-Toth, M. and Ferec, C. 2007. Signal peptide variants that impair secretion of pancreatic secretory trypsin inhibitor (SPINK1) cause autosomal dominant hereditary pancreatitis. *Hum Mutat*, **28**(5):469-476.

Klee, E. W. and Ellis, L. B. M. 2005. Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, **6**:256.

Klug, G., Jager, A., Heck, C. and Rauhut, R. 1997. Identification, sequence analysis, and expression of the *lepB* gene for a leader peptidase in *Rhodobacter capsulatus*. *Mol Gen Genet*, **253**(6):666-673.

Kochl, R., Alken, M., Rutz, C., Krause, G., Oksche, A., Rosenthal, W. and Schulein, R. 2002. The signal peptide of the G protein-coupled human endothelin B receptor is necessary for translocation of the N-terminal tail across the endoplasmic reticulum membrane. *J Biol Chem*, **277**(18):16131-16138.

Koike, A., Niwa, Y. and Takagi, T. 2005. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, **21**(7):1227-1236.

Koren, R., Burstein, Y. and Soreq, H. 1983. Synthetic leader peptide modulates secretion of proteins from microinjected *Xenopus* oocytes. *Proc Natl Acad Sci U S A*, **80**(23):7205-7209.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M. P., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. 2007. EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res*, **35**(Database issue):D16-20.

Kurys, G., Tagaya, Y., Bamford, R., Hanover, J. A. and Waldmann, T. A. 2000. The long signal peptide isoform and its alternative processing direct the intracellular trafficking of interleukin-15. *J Biol Chem*, **275**(39):30653-30659.

Kutz, W. E., Wang, L. W., Dagoneau, N., Odrcic, K. J., Cormier-Daire, V., Traboulsi, E. I. and Apte, S. S. 2008. Functional analysis of an ADAMTS10 signal peptide mutation in Weill-Marchesani syndrome demonstrates a long-range effect on secretion of the full-length enzyme. *Hum Mutat*, **29**(12):1425-1434.

Kyte, J. and Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**(1):105-132.

Ladunga, I. 2000. Large-scale predictions of secretory proteins from mammalian genomic and EST sequences. *Curr Opin Biotechnol*, **11**(1):13-18.

Laforet, G. A. and Kendall, D. A. 1991. Functional limits of conformation, hydrophobicity, and steric constraints in prokaryotic signal peptide cleavage regions. Wild type transport by a simple polymeric signal sequence. *J Biol Chem*, **266**(2):1326-1334.

Lal, P., Au-Young, J., Reddy, R., Murry, L. E. and Mathur, P. 1999. Signal peptide-containing proteins. In. Edited by WIPO. USA.

Lam, S. L., Kirby, S. and Schryvers, A. B. 2003. Foreign signal peptides can constitute a barrier to functional expression of periplasmic proteins in *Haemophilus influenzae*. *Microbiology*, **149**(Pt 11):3155-3164.

Lammertyn, E. and Anne, J. 1998. Modifications of *Streptomyces* signal peptides and their effects on protein production and secretion. *FEMS Microbiol Lett*, **160**(1):1-10.

Langer, R. 1998. Drug delivery and targeting. *Nature*, **392**(6679 Suppl):5-10.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**(5131):208-214.

Le Loir, Y., Nouaille, S., Commissaire, J., Bretigny, L., Gruss, A. and Langella, P. 2001. Signal peptide and propeptide optimization for heterologous protein secretion in *Lactococcus lactis*. *Appl Environ Microbiol*, **67**(9):4119-4127.

Le Loir, Y., Azevedo, V., Oliveira, S. C., Freitas, D. A., Miyoshi, A., Bermudez-Humaran, L. G., Nouaille, S., Ribeiro, L. A., Leclercq, S., Gabriel, J. E., Guimaraes, V. D., Oliveira, M. N., Charlier, C., Gautier, M. and Langella, P. 2005. Protein secretion in *Lactococcus lactis* : an efficient way to increase the overall heterologous protein production. *Microb Cell Fact*, **4**(1):2.

Lee, N., Goodlett, D. R., Ishitani, A., Marquardt, H. and Geraghty, D. E. 1998. HLA-E surface expression depends on binding of TAP-dependent peptides derived from certain HLA class I signal sequences. *J Immunol,* **160**(10):4951-4960.

Lemberg, M. K., Bland, F. A., Weihofen, A., Braud, V. M. and Martoglio, B. 2001. Intramembrane proteolysis of signal peptides: an essential step in the generation of HLA-E epitopes. *J Immunol*, **167**(11):6441-6446.

Lemberg, M. K. and Martoglio, B. 2002. Requirements for signal peptide peptidase-catalyzed intramembrane proteolysis. *Mol Cell*, **10**(4):735-744.

Levine, C. G., Mitra, D., Sharma, A., Smith, C. L. and Hegde, R. S. 2005. The efficiency of protein compartmentalization into the secretory pathway. *Mol Biol Cell*, **16**(1):279-291.

Li, P., Beckwith, J. and Inouye, H. 1988. Alteration of the amino terminus of the mature sequence of a periplasmic protein can severely affect protein export in *Escherichia coli*. *Proc Natl Acad Sci U S A*, **85**(20):7685-7689.

Li, Y., Luo, L., Thomas, D. Y. and Kang, C. Y. 1994. Control of expression, glycosylation, and secretion of HIV-1 gp120 by homologous and heterologous signal sequences. *Virology*, **204**(1):266-278.

Li, Y., Bergeron, J. J., Luo, L., Ou, W. J., Thomas, D. Y. and Kang, C. Y. 1996. Effects of inefficient cleavage of the signal sequence of HIV-1 gp 120 on its association with calnexin, folding, and intracellular transport. *Proc Natl Acad Sci U S A*, **93**(18):9606-9611.

Li, W. and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13):1658-1659.

Li, Y., Wen, Z., Zhou, C., Tan, F. and Li, M. 2008. Effects of neighboring sequence environment in predicting cleavage sites of signal peptides. *Peptides*, **29**(9):1498-1504.

Lindemann, D., Pietschmann, T., Picard-Maureau, M., Berg, A., Heinkelein, M., Thurow, J., Knaus, P., Zentgraf, H. and Rethwilm, A. 2001. A particle-associated glycoprotein signal peptide essential for virus maturation and infectivity. *J Virol*, **75**(13):5762-5771.

Lingappa, V. R., Lingappa, J. R. and Blobel, G. 1979. Chicken ovalbumin contains an internal signal sequence. *Nature*, **281**(5727):117-121.

Liu, L., Li, J., Tian, X., Ren, D. and Lin, J. 2005. Information theory in prediction of cleavage sites of signal peptides. *Protein and Peptides Letters*, **12**:339-342.

Liu, D. Q., Liu, H., Shen, H. B., Yang, J. and Chou, K. C. 2007. Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. *Amino Acids*, **32**(4):493-496.

Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, L. and Darnell, J. 2004. Translocation of secretory proteins across the ER membrane. In: *Molecular Cell Biology.* 5th edn, pp. 659-666. ISBN-10: 0716743663.

Long, E. O. 1998. Signal sequences stop killer cells. *Nature*, **391**(6669):740-741, 743.

Louie, B., Tarczy-Hornoch, P., Higdon, R. and Kolker, E. 2008. Validating annotations for uncharacterized proteins in *Shewanella oneidensis*. *Omics*, **12**(3):211-215.

Luo, C., Roussel, P., Dreier, J., Page, M. G., Paetzel, M. 2009. Crystallographic analysis of bacterial signal peptidase in ternary complex with arylomycin A2 and a β-Sultam inhibitor. *Biochemistry*, **48**(38): 8976-8984.

Lyko, F., Martoglio, B., Jungnickel, B., Rapoport, T. A. and Dobberstein, B. 1995. Signal sequence processing in rough microsomes. *J Biol Chem*, **270**(34):19873-19878.

Maetschke, S., Towsey, M. and Bodén, M. 2005. BLOMAP: an encoding of amino acids which improves signal peptide cleavage site prediction. In: Chen, Y. P. P., Wong, L. S. (Eds), *Proceedings of the 3rd Asia-Pacific Bioinformatics Conference Singapore*: Imperial College Press, pp. 141-150.

Marra, M. A., Jones, S. J., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y., Cloutier, A., Coughlin, S. M., Freeman, D., Girn, N., Griffith, O. L., Leach, S. R., Mayo, M., McDonald, H., Montgomery, S. B., Pandoh, P. K., Petrescu, A. S., Robertson, A. G., Schein, J. E., Siddiqui, A., Smailus, D. E., Stott, J. M., Yang, G. S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T. F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G. A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R. C., Krajden, M., Petric, M., Skowronski, D. M., Upton, C. and Roper, R. L. 2003. The Genome sequence of the SARS-associated coronavirus. *Science*, **300**(5624):1399-1404.

Martoglio, B., Graf, R. and Dobberstein, B. 1997. Signal peptide fragments of preprolactin and HIV-1 p-gp160 interact with calmodulin. *Embo J*, **16**(22):6636-6645.

Martoglio, B. and Dobberstein, B. 1998. Signal sequences: more than just greasy peptides. *Trends Cell Biol*, **8**(10):410-415.

Martoglio, B. 2003a. Intramembrane proteolysis and post-targeting functions of signal peptides. *Biochem Soc Trans*, **31**(Pt 6):1243-1247.

Martoglio, B. and Golde, T. E. 2003b. Intramembrane-cleaving aspartic proteases and disease: presenilins, signal peptide peptidase and their homologs. *Hum Mol Genet*, **12**(2):R201-206.

Marzi, A., Akhavan, A., Simmons, G., Gramberg, T., Hofmann, H., Bates, P., Lingappa, V. R. and Pohlmann, S. 2006. The signal peptide of the ebolavirus glycoprotein influences interaction with the cellular lectins DC-SIGN and DC-SIGNR. *J Virol*, **80**(13):6305-6317.

Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, **405**(2):442-451.

Matoba, S. and Ogrydziak, D. M. 1998. Another factor besides hydrophobicity can affect signal peptide interaction with signal recognition particle. *J Biol Chem*, **273**(30):18841-18847.

Mayo, M. *2005* Bayesian sequence learning for predicting protein cleavage points. In: Ho, T. B. Cheung, D. Liu, H. (Eds), *Advances in Knowledge Discovery and Data Mining, 9th Pacific-Asia Conference, PAKDD*, Hanoi, Vietnam: Springer, pp. 192-202.

McDonald, I. K. and Thornton, J. M. 1994. Satisfying hydrogen bonding potential in proteins. *J Mol Biol*, **238**(5):777-793.

McKnight, C. J., Stradley, S. J., Jones, J. D. and Gierasch, L. M. 1991. Conformational and membrane-binding properties of a signal sequence are largely unaltered by its adjacent mature region. *Proc Natl Acad Sci U S A*, **88**(13):5799-5803.

McLauchlan, J., Lemberg, M. K., Hope, G. and Martoglio, B. 2002. Intramembrane proteolysis promotes trafficking of hepatitis C virus core protein to lipid droplets. *Embo J*, **21**(15):3980-3988.

Menne, K. M., Hermjakob, H. and Apweiler, R. 2000. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, **16**(8):741-742.
Dataset download: ftp://ftp.ebi.ac.uk/pub/contrib/swissprot/testsets/signal

Metropolis, N. A., Rosenbluth, A. W., Rosenbluth, N. M., Teller, A. H. and Teller, E. 1953. Equation of State Calculations by Fast Computing Machines. *J Chem Phys*, **21**(6):1087-1092.

Milstein, C., Brownlee, G. G., Harrison, T. M. and Mathews, M. B. 1972. A possible precursor of immunoglobulin light chains. *Nat New Biol*, **239**(91):117-120.

Molhoj, M. and Degan, F. D. 2004. Leader sequences are not signal peptides. *Nat Biotechnol*, **22**(12):1502.

Mothes, W., Jungnickel, B., Brunner, J. and Rapoport, T. A. 1998. Signal sequence recognition in cotranslational translocation by protein components of the endoplasmic reticulum membrane. *J Cell Biol*, **142**(2):355-364.

Mount, D. W. 2001. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, 1st edn. ISBN-10: 0879696087.

Mukherjee, N. and Mukherjee, S. 2002. Predicting signal peptides with support vector machines. In: Lee, S. W. and Verri, A *(Eds), Pattern Recognition with Support Vector Machines*, vol. 2388/2002. Springer Berlin / Heidelberg, pp. 487-500.

Musial-Siwek, M., Yeagle, P. L. and Kendall, D. A. 2008. A small subset of signal peptidase residues are perturbed by signal peptide binding. *Chem Biol Drug Des*, **72**(2):140-146.

Nagaraj, S. H., Gasser, R. B., Ranganathan, S. 2008. Needles in the EST haystack: large-scale identification and analysis of excretory-secretory (ES) proteins in parasitic nematodes using expressed sequence tags (ESTs). *PLoS Negl Trop Dis*, **2**(9):e301.

Nakashima, H. and Nishikawa, K. 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J Mol Biol*, **238**(1):54-61.

Needleman, S. B. and Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**(3):443-453.

Nene, V. and Bishop, R. 2001. Trapping parasite secretory proteins in baker's yeast. *Trends Parasitol*, **17**(9):407-409.

Ng, D. T., Brown, J. D. and Walter, P. 1996. Signal sequences specify the targeting route to the endoplasmic reticulum membrane. *J Cell Biol*, **134**(2):269-278.

Ng, S. Y., Chaban, B., VanDyke, D. J. and Jarrell, K. F. 2007. Archaeal signal peptidases. *Microbiology*, **153**(Pt 2):305-314.

Nickel, W. and Rabouille, C. 2009. Mechanisms of regulated unconventional protein secretion. *Nat Rev Mol Cell Biol*, **10**(2):148-155.

Nielsen, H., Engelbrecht, J., von Heijne, G. and Brunak, S. 1996. Defining a similarity threshold for a functional protein sequence pattern: the signal peptide cleavage site. *Proteins*, **24**(2):165-177.

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

Nielsen, H. and Krogh, A. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. In: Glasglow, J., *et al.* (Eds), *Proc of the Sixth Int Conf Intell Syst Mol Biol*, Menlo Park, California, AAAI Press, **6**:122-130.

Nilsson, I. and von Heijne, G. 1992. A signal peptide with a proline next to the cleavage site inhibits leader peptidase when present in a *sec*-independent protein. *FEBS Lett*, **299**(3):243-246.

Nouaille, S., Bermudez-Humaran, L. G., Adel-Patient, K., Commissaire, J., Gruss, A., Wal, J. M., Azevedo, V., Langella, P. and Chatel, J. M. 2005. Improvement of bovine beta-lactoglobulin production and secretion by *Lactococcus lactis*. *Braz J Med Biol Res*, **38**(3):353-359.

Novak, P. and Dev, I. K. 1988. Degradation of a signal peptide by protease IV and oligopeptidase A. *J Bacteriol*, **170**(11):5067-5075.

O'Callaghan, C. A., Tormo, J., Willcox, B. E., Braud, V. M., Jakobsen, B. K., Stuart, D. I., McMichael, A. J., Bell, J. I. and Jones, E. Y. 1998. Structural features impose tight peptide binding specificity in the nonclassical MHC molecule HLA-E. *Mol Cell*, **1**(4):531-541.

Olczak, M. and Olczak, T. 2006. Comparison of different signal peptides for protein secretion in nonlytic insect cell system. *Anal Biochem*, **359**(1):45-53.

Oliver, D. 1985. Protein secretion in *Escherichia coli*. *Annu Rev Microbiol*, **39**:615-648.

Olivera, B. M., Walker, C., Cartier, G. E., Hooper, D., Santos, A. D., Schoenfeld, R., Shetty, R., Watkins, M., Bandyopadhyay, P. and Hillyard, D. R. 1999. Speciation of cone snails and interspecific hyperdivergence of their venom peptides. Potential evolutionary significance of introns. *Ann N Y Acad Sci*, **870**:223-237.

Osborne, R. S. and Silhavy, T. J. 1993. PrlA suppressor mutations cluster in regions corresponding to three distinct topological domains. *Embo J*, **12**(9):3391-3398.

Ouzzine, M., Magdalou, J., Burchell, B. and Fournel-Gigleux, S. 1999. Expression of a functionally active human hepatic UDP-glucuronosyltransferase (UGT1A6) lacking the N-terminal signal sequence in the endoplasmic reticulum. *FEBS Lett*, **454**(3):187-191.

Oxender, D. L., Anderson, J. J., Daniels, C. J., Landick, R., Gunsalus, R. P., Zurawski, G. and Yanofsky, C. 1980. Amino-terminal sequence and processing of the precursor of the leucine-specific binding protein, and evidence for conformational differences between the precursor and the mature form. *Proc Natl Acad Sci U S A*, **77**(4):2005-2009.

Panchenko, A. R. and Bryant, S. H. 2002. A comparison of position-specific score matrices based on sequence and structure alignments. *Protein Sci*, **11**(2):361-370.

Paetzel, M., Dalbey, R. E. and Strynadka, N. C. 1998. Crystal structure of a bacterial signal peptidase in complex with a beta-lactam inhibitor. *Nature*, **396**(6707):186-190.

Paetzel, M., Dalbey, R. E. and Strynadka, N. C. 2000. The structure and mechanism of bacterial type I signal peptidases. A novel antibiotic target. *Pharmacol Ther*, **87**(1):27-49.

Paetzel, M. and Strynadka, N. C. 2001. Signal peptide cleavage in the *E. coli* membrane. *CSBMCB Bulletin*.

Paetzel, M., Dalbey, R. E. and Strynadka, N. C. 2002a. Crystal structure of a bacterial signal peptidase apoenzyme: implications for signal peptide binding and the Ser-Lys dyad mechanism. *J Biol Chem*, **277**(11):9512-9519.

Paetzel, M., Karla, A., Strynadka, N. C. and Dalbey, R. E. 2002b. Signal peptidases. *Chem Rev*, **102**(12):4549-4580.

Paetzel, M., Goodall, J. J., Kania, M., Dalbey, R. E. and Page, M. G. 2004. Crystallographic and biophysical analysis of a bacterial signal peptidase in complex with a lipopeptide-based inhibitor. *J Biol Chem*, **279**(29):30781-30790.

Palazzo, A. F., Springer, M., Shibata, Y., Lee, C. S., Dias, A. P. and Rapoport, T. A. 2007. The signal sequence coding region promotes nuclear export of mRNA. *PLoS Biol*, **5**(12):e322.

Park, S., Liu, G., Topping, T. B., Cover, W. H. and Randall, L. L. 1988. Modulation of folding pathways of exported proteins by the leader sequence. *Science*, **239**(4843):1033-1035.

Pascarella, S. and Bossa, F. 1989. CLEAVAGE: a microcomputer program for predicting signal sequence cleavage sites. *Comput Appl Biosci*, **5**(1):53-54.

Pennisi, E. 1999. Keeping genome databases clean and up to date. *Science*, **286**(5439):447-450.

Perna, N. T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J., Kirkpatrick, H. A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E. J., Davis, N. W., Lim, A., Dimalanta, E. T., Potamousis, K. D., Apodaca, J., Anantharaman, T. S., Lin, J., Yen, G., Schwartz, D. C., Welch, R. A. and Blattner, F. R. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**(6819):529-533.

Péterfy, M, Gyuris, T, Takács, L. 2000. Signal-exon trap: a novel method for the identification of signal sequences from genomic DNA. *Nucleic Acids Res*, 28(7):E26.

Pfanner, N., Hartl, F. U. and Neupert, W. 1988. Import of proteins into mitochondria: a multi-step process. *Eur J Biochem*, **175**(2):205-212.

Pfeiffer, T., Pisch, T., Devitt, G., Holtkotte, D. and Bosch, V. 2006. Effects of signal peptide exchange on HIV-1 glycoprotein expression and viral infectivity in mammalian cells. *FEBS Lett*, **580**(15):3775-3778.

Pidasheva, S., Canaff, L., Simonds, W. F., Marx, S. J. and Hendy, G. N. 2005. Impaired cotranslational processing of the calcium-sensing receptor due to signal peptide missense mutations in familial hypocalciuric hypercalcemia. *Hum Mol Genet*, **14**(12):1679-1690.

Plath, K., Mothes, W., Wilkinson, B. M., Stirling, C. J. and Rapoport, T. A. 1998. Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell*, **94**(6):795-807.

Plewczynski, D., Slabinski, L., Ginalski, K. and Rychlewski, L. 2008. Prediction of signal peptides in protein sequences by neural networks. *Acta Biochim Pol*, **55**(2):261-267.

Pool, M. R. 2005. Signal recognition particles in chloroplasts, bacteria, yeast and mammals (review). *Mol Membr Biol*, **22**(1-2):3-15.

Popowicz, A. M. and Dash, P. F. 1988. SIGSEQ: a computer program for predicting signal sequence cleavage sites. *Comput Appl Biosci*, **4**(3):405-406.

Pradel, N., Ye, C. and Wu, L. F. 2004. A cleavable signal peptide is required for the full function of the polytopic inner membrane protein FliP of *Escherichia coli*. *Biochem Biophys Res Commun*, **319**(4):1276-1280.

Prinz, W. A., Spiess, C., Ehrmann, M., Schierle, C. and Beckwith, J. 1996. Targeting of signal sequenceless proteins for export in *Escherichia coli* with altered protein translocase. *EMBO J*, **15**(19):5209-5217.

Prudovsky, I., Mandinova, A., Soldi, R., Bagala, C., Graziani, I., Landriscina, M., Tarantini, F., Duarte, M., Bellum, S., Doherty, H. and Maciag, T. 2003. The non-classical export routes: FGF1 and IL-1alpha point the way. *J Cell Sci*, **116**(Pt 24):4871-4881.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**(Database issue):D501-504.

Pugsley, A. P. 1989. Protein targeting. Academic Press, 1st edn. ISBN-10: 0125667701.

Purdue, P. E., Allsop, J., Isaya, G., Rosenberg, L. E. and Danpure, C. J. 1991. Mistargeting of peroxisomal L-alanine:glyoxylate aminotransferase to mitochondria in primary hyperoxaluria patients depends upon activation of a cryptic mitochondrial targeting sequence by a point mutation. *Proc Natl Acad Sci U S A*, **88**(23):10900-10904.

Rabiner, L. R. *1989* A tutorial on hidden Markov models and selected applications in speech recognition. In: *IEEE*. **77**(72): 257-286.

Racchi, M., Watzke, H.H., High, K.A. and Lively, M.O. (1993) Human coagulation factor X deficiency caused by a mutant signal peptide that blocks cleavage by signal peptidase but not targeting and translocation to the endoplasmic reticulum. *J Biol Chem*, **268**, 5735–5740.

Rajalahti, T., Huang, F., Klement, M. R., Pisareva, T., Edman, M., Sjostrom, M., Wieslander, A. and Norling, B. 2007. Proteins in different Synechocystis compartments have distinguishing N-terminal features: a combined proteomics and multivariate sequence analysis. *J Proteome Res*, **6**(7):2420-2434.

Rajpar, M. H., Koch, M. J., Davies, R. M., Mellody, K. T., Kielty, C. M. and Dixon, M. J. 2002. Mutation of the signal peptide region of the bicistronic gene DSPP affects translocation to the endoplasmic reticulum and results in defective dentine biomineralization. *Hum Mol Genet*, **11**(21):2559-2565.

Rapoport, T. A., Jungnickel, B. and Kutay, U. 1996. Protein transport across the eukaryotic endoplasmic reticulum and bacterial inner membranes. *Annu Rev Biochem*, **65**:271-303.

Ravn, P., Arnau, J., Madsen, S. M., Vrang, A. and Israelsen, H. 2003. Optimization of signal peptide SP310 for heterologous protein production in *Lactococcus lactis*. *Microbiology*, **149**(8):2193-201.

Reczko, M., Fiziev, P., Staub, E. and Hatzigeorgiou, A. 2002. Finding signal peptides in human protein sequences using recurrent neural networks. In: Guigó, R., Gusfield, D. (Eds), *Algorithms in Bioinformatics.* vol. 2452/2002. Springer-Verlag, pp. 60-67.

Reed, J. L., Famili, I., Thiele, I. and Palsson, B. O. 2006. Towards multidimensional genome annotation. *Nat Rev Genet*, **7**(2):130-141.

Rehm, A., Stern, P., Ploegh, H. L. and Tortorella, D. 2001. Signal peptide cleavage of a type I membrane protein, HCMV US11, is dependent on its membrane anchor. *Embo J*, **20**(7):1573-1582.

Reynolds, S. M., Käll, L., Riffle, M. E., Bilmes, J. A. and Nobel, W. S. 2008. Transmembrane topology and signal peptide prediction using dynamic Bayesian networks. *PLoS Comput Biol*, **4**(11):e1000213.

Rhodes, G. 2006. Crystallography made crystal clear: a guide for users of macromolecular models. Academic Press, 3rd edn. ISBN-10: 0125870728.

Rice, P., Longden, I. and Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, **16**(6):276-277.

Rittig, S., Siggaard, C., Ozata, M., Yetkin, I., Gregersen, N., Pedersen, E. B. and Robertson, G. L. 2002. Autosomal dominant neurohypophyseal diabetes insipidus due to substitution of histidine for tyrosine(2) in the vasopressin moiety of the hormone precursor. *J Clin Endocrinol Metab*, **87**(7):3351-3355.

Robakis, T., Bak, B., Lin, S. H., Bernard, D. J. and Scheiffele, P. 2008. An internal signal sequence directs intramembrane proteolysis of a cellular immunoglobulin domain protein. *J Biol Chem*, **283**(52):36369-36376.

Roggenkamp, R., Dargatz, H. and Hollenberg, C. P. 1985. Precursor of beta-lactamase is enzymatically inactive. Accumulation of the preprotein in *Saccharomyces cerevisiae*. *J Biol Chem*, **260**(3):1508-1512.

Romisch, K. 1999. Surfing the *Sec61* channel: bidirectional protein translocation across the ER membrane. *J Cell Sci*, **112** (Pt 23):4185-4191.

Ronald, L. S., Yakovenko, O., Yazvenko, N., Chattopadhyay, S., Aprikian, P., Thomas, W. E. and Sokurenko, E. V. 2008. Adaptive mutations in signal peptide of the type 1 fimbrial adhesin of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A*, **105**(31):10937-42.

Rosander, A., Bjerketorp, J., Frykberg, L. and Jacobsson, K. 2002. Phage display as a novel screening method to identify extracellular proteins. *J Microbiol Methods*, **51**(1):43-55.

Rusch, S. L., Chen, H., Izard, J. W. and Kendall, D. A. 1994. Signal peptide hydrophobicity is finely tailored for function. *J Cell Biochem*, **55**(2):209-217.

Rusch, S. L., Mascolo, C. L., Kebir, M. O. and Kendall, D. A. 2002. Juxtaposition of signal-peptide charge and core region hydrophobicity is critical for functional signal peptides. *Arch Microbiol*, **178**:306-310.

Russel, M. and Model, P. 1981. A mutation downstream from the signal peptidase cleavage site affects cleavage but not membrane insertion of phage coat protein. *Proc Natl Acad Sci U S A*, **78**(3):1717-1721.

Rutkowski, D. T., Lingappa, V. R. and Hegde, R. S. 2001. Substrate-specific regulation of the ribosome- translocon junction by N-terminal signal sequences. *Proc Natl Acad Sci U S A*, **98**(14):7823-7828.

Rutkowski, D. T., Ott, C. M., Polansky, J. R. and Lingappa, V. R. 2003. Signal sequences initiate the pathway of maturation in the endoplasmic reticulum lumen. *J Biol Chem*, **278**(32):30365-30372.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. http://www.R-project.org

Sacksteder, K. A. and Gould, S. J. 2000. The genetics of peroxisome biogenesis. *Annu Rev Genet*, **34**:623-652.

Sali, A. and Blundell, T. L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**(3):779-815.

Santini, J. T., Jr., Cima, M. J. and Langer, R. 1999. A controlled-release microchip. *Nature*, **397**(6717):335-338.

Schaaf, A., Tintelnot, S., Baur, A., Reski, R., Gorr, G. and Decker, E. L. 2005. Use of endogenous signal sequences for transient production and efficient secretion by moss (*Physcomitrella patens*) cells. *BMC Biotechnol*, **5**:30.

Schartl, M., Wilde, B. and Hornung, U. 1998. Triplet repeat variability in the signal peptide sequence of the *Xmrk* receptor tyrosine kinase gene in *Xiphophorus* fish. *Gene*, **224**(1-2):17-21.

Schatz, G. 1993. The protein import machinery of mitochondria. *Protein Sci*, **2**(2):141-146.

Schneider, G. and Fechner, U. 2004. Advances in the prediction of protein targeting signals. *Proteomics*, **4**(6):1571-1580.

Scott, M., Lu, G., Hallett, M. and Thomas, D. Y. 2004. The Hera database and its use in the characterization of endoplasmic reticulum proteins. *Bioinformatics*, 20(6):937-944.

Serruto, D. and Galeotti, C. L. 2004. The signal peptide sequence of a lytic transglycosylase of *Neisseria meningitidis* is involved in regulation of gene expression. *Microbiology*, **150**(Pt 5):1427-1437.

Serruto, D., Adu-Bobie, J., Capecchi, B., Rappuoli, R., Pizza, M. and Masignani, V. 2004. Biotechnology and vaccines: application of functional genomics to *Neisseria meningitidis* and other bacterial pathogens. *J Biotechnol*, **113**(1-3):15-32.

Shen, M. Y. and Sali, A. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, **15**:2507-2524.

Shen, H. B. and Chou, K. C. 2007. Signal-3L: A 3-layer approach for predicting signal peptides. *Biochem Biophys Res Commun*, **363**(2):297-303.

Sidhu, A. and Yang, Z. R. 2006. Prediction of signal peptides using bio-basis function neural networks and decision trees. *Appl Bioinformatics*, **5**(1):13-19.

Simon, S. M., Peskin, C. S. and Oster, G. F. 1992. What drives the translocation of proteins? *Proc Natl Acad Sci U S A*, **89**(9):3770-3774.

Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**(20):3940-3941.

Skach, W. R. 2007. The expanding role of the ER translocon in membrane protein folding. *J Cell Biol*, **179**(7):1333-1335.

Small, I., Peeters, N., Legeai, F. and Lurin, C. 2004. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**(6):1581-1590.

Spiess, M. 1995. Heads or tails--what determines the orientation of proteins in the membrane. *FEBS Lett*, **369**(1):76-79.

Stamnes, M. A., Shieh, B. H., Chuman, L., Harris, G. L. and Zuker, C. S. 1991. The cyclophilin homolog ninaA is a tissue-specific integral membrane protein required for the proper synthesis of a subset of *Drosophila rhodopsins*. *Cell*, **65**(2):219-227.

Summers, R. G. and Knowles, J. R. 1989. Illicit secretion of a cytoplasmic protein into the periplasm of *Escherichia coli* requires a signal peptide plus a portion of the cognate secreted protein. Demarcation of the critical region of the mature protein. *J Biol Chem*, **264**(33):20074-20081.

Summers, R. G., Harris, C. R. and Knowles, J. R. 1989. A conservative amino acid substitution, arginine for lysine, abolishes export of a hybrid protein in *Escherichia coli*. Implications for the mechanism of protein secretion. *J Biol Chem*, **264**(33):20082-20088.

Sun, J. J. and Wang, L. 2008 Predicting signal peptides and their cleavage sites using support vector machines and improved position weight matrices. In: *Proceedings of the 4th International Conference on Natural Computation*: ICNC, **5**:95-99.

Swanton, E. and High, S. 2006. ER targeting signals: more than meets the eye? *Cell*, **127**(5):877-879.

Sweet, R. M. and Eisenberg, D. 1983. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J Mol Biol*, **171**(4):479-488.

Symoens, S., Malfait, F., Renard, M., Andre, J., Hausser, I., Loeys, B., Coucke, P. and De Paepe, A. 2008. COL5A1 signal peptide mutations interfere with protein secretion and cause classic Ehlers-Danlos syndrome. *Hum Mutat*, **30**(2):E395-E403.

Szabady, R. L., Peterson, J. H., Skillman, K. M. and Bernstein, H. D. 2005. An unusual signal peptide facilitates late steps in the biogenesis of a bacterial autotransporter. *Proc Natl Acad Sci U S A*, **102**(1):221-226.

Tabe, L., Krieg, P., Strachan, R., Jackson, D., Wallis, E. and Colman, A. 1984. Segregation of mutant ovalbumins and ovalbumin-globin fusion proteins in *Xenopus* oocytes. Identification of an ovalbumin signal sequence. *J Mol Biol*, **180**(3):645-666.

Tan, N. S., Ho, B. and Ding, J. L. 2002. Engineering a novel secretion signal for cross-host recombinant protein expression. *Protein Eng*, **15**(4):337-345.

Tan, T. W., Choo, K. H., Tong, J. C., Tammi, M. T. and Bajic, V. 2005. Biological databases and web services: metrics for qualitative analysis. In: Bajic, V. and Tan, T. W. (Eds), *Information Processing and Living Systems*, vol. 2. World Scientific Publishing Co., 1st edn, pp. 771-778. ISBN-10: 1860945635.

Taylor, P. D., Toseland, C. P., Attwood, T. K. and Flower, D. R. 2006. LIPPRED: A web server for accurate prediction of lipoprotein signal sequences and cleavage sites. *Bioinformation*, **1**(5):176-179.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. 2000. Automatic extraction of protein interactions from scientific abstracts. *Pac Symp Biocomput*:541-552.

Thornton, J., Blakey, D., Scanlon, E. and Merrick, M. 2006. The ammonia channel protein AmtB from *Escherichia coli* is a polytopic membrane protein with a cleavable signal peptide. *FEMS Microbiol Lett*, **258**(1):114-120.

Tjalsma, H., Bolhuis, A., van Roosmalen, M. L., Wiegert, T., Schumann, W., Broekuizen, C. P., Quax, W. J., Venema, G., Bron, S., van Dijl, J. M. 1998. Functional analysis of the secretory precursor processing machinery of *Bacillus subtilis*: identification of a eubacterial homolog of archaeal and eukaryotic signal peptidases. *Genes Dev*, **12**:2318–2331.

Tjalsma, H., Kontinen, V. P., Pragai, Z., Wu, H., Meima, R., Venema, G., Bron, S., Sarvas, M. and van Dijl, J. M. 1999. The role of lipoprotein processing by signal peptidase II in the Gram-positive eubacterium *Bacillus subtilis*. Signal peptidase II is required for the efficient secretion of alpha-amylase, a non-lipoprotein. *J Biol Chem*, **274**(3):1698-1707.

Tong, J. C., Tan, T. W. and Ranganathan, S. 2004. Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci*, **13**(9):2523-2532.

Totrov, M. and Abagyan, R. 2001. Rapid boundary element solvation electrostatics calculations in folding simulations: successful folding of a 23-residue peptide. *Biopolymers*, **60**(2):124-133.

Tsuchiya, Y., Morioka, K., Shirai, J., Yokomizo, Y. and Yoshida, K. 2003. Gene design of signal sequence for effective secretion of protein. *Nucleic Acids Res Suppl*(3):261-262.

Tsujibo, H., Fujimoto, K., Tanno, H., Miyamoto, K., Imada, C., Okami, Y. and Inamori, Y. 1994. Gene sequence, purification and characterization of N-acetyl-beta-glucosaminidase from a marine bacterium, *Alteromonas sp*. strain O-7. *Gene*, **146**(1):111-115.

Tuteja, R. 2005. Type I signal peptidase: an overview. *Arch Biochem Biophys*, **441**(2):107-111.

van Roosmalen, M. L., Geukens, N., Jongbloed, J. D., Tjalsma, H., Dubois, J. Y., Bron, S., van Dijl, J. M. and Anne, J. 2004. Type I signal peptidases of Gram-positive bacteria. *Biochim Biophys Acta*, **1694**(1-3):279-297.

van Vliet, C., Thomas, E. C., Merino-Trigo, A., Teasdale, R. D. and Gleeson, P. A. 2003. Intracellular sorting and transport of proteins. *Prog Biophys Mol Biol*, **83**(1):1-45.

van Voorst, F. and De Kruijff, B. 2000. Role of lipids in the translocation of proteins across membranes. *Biochem J*, **347 Pt 3**:601-612.

Vapnik, V.N. 1998. Statistical learning theory. Wiley-Interscience, 1st edn. ISBN-10: 0471030031.

Vert, J. P. 2002. Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings. *Pac Symp Biocomput*, **7**:649-660.

Viklund, H., Bernsel, A., Skwark, M. and Elofsson, A. 2008. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics App Notes*, **24**(24):2928-2929.

von Heijne, G. 1982. Signal sequences are not uniformly hydrophobic. *J Mol Biol*, **159**(3):537-541.

von Heijne, G. 1983. Patterns of amino acids near signal-sequence cleavage sites. *Eur J Biochem*, **133**(1):17-21.

von Heijne, G. 1984a. How signal sequences maintain cleavage specificity. *J Mol Biol*, **173**(2):243-251.

von Heijne, G. 1984b. Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implications for protein export in prokaryotic and eukaryotic cells. *Embo J*, **3**(10):2315-2318.

von Heijne, G. 1985. Signal sequences. The limits of variation. *J Mol Biol*, **184**(1):99-105.

von Heijne, G. 1986a. A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res*, **14**(11):4683-4690.

von Heijne, G. 1986b. Net N-C charge imbalance may be important for signal sequence function in bacteria. *J Mol Biol*, **192**(2):287-290.

von Heijne, G. and Abrahmsen, L. 1989. Species-specific variation in signal peptide design. Implications for protein secretion in foreign hosts. *FEBS Lett*, **244**(2):439-446.

von Heijne, G. 1990. The signal peptide. *J Membr Biol*, **115**(3):195-201.

von Heijne, G. 1994. Design of protein targeting signals and membrane protein engineering. In: Wrede, P. and Schneider, G. (Eds), *Concepts in Protein Engineering and Design: An Introduction.* Walter de Gruyter, Inc., 1st edn, pp. 263-279. ISBN-10: 3110129752.

von Heijne, G. 1998. Life and death of a signal peptide. *Nature*, **396**(6707):111, 113.

Walker, J. M. 2005. The proteomics protocols handbook. Humana Press, 1st edn. ISBN-10: 1588295931.

Wall, L. 2000. Programming Perl. O'Reilly Media, Inc., 3rd edn. ISBN-10: 0596000278.

Walter, P. and Blobel, G. 1980. Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum. *Proc Natl Acad Sci U S A*, **77**(12):7112-7116.

Walter, P., Ibrahimi, I. and Blobel, G. 1981a. Translocation of proteins across the endoplasmic reticulum. I. Signal recognition protein (SRP) binds to *in vitro*-assembled polysomes synthesizing secretory protein. *J Cell Biol*, **91**(2 Pt 1):545-550.

Walter, P. and Blobel, G. 1981b. Translocation of proteins across the endoplasmic reticulum. II. Signal recognition protein (SRP) mediates the selective binding to microsomal membranes of *in vitro*-assembled polysomes synthesizing secretory protein. *J Cell Biol*, **91**(2 Pt 1):551-556.

Walter, P. and Blobel, G. 1981c. Translocation of proteins across the endoplasmic reticulum III. Signal recognition protein (SRP) causes signal sequence-dependent and site-specific arrest of chain elongation that is released by microsomal membranes. *J Cell Biol*, **91**(2 Pt 1):557-561.

Walter, P. and Blobel, G. 1982. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, **299**(5885):691-698.

Walter, P., Gilmore, R. and Blobel, G. 1984. Protein translocation across the endoplasmic reticulum. *Cell*, **38**(1):5-8.

Walter, P. and Lingappa, V. R. 1986. Mechanism of protein translocation across the endoplasmic reticulum membrane. *Annu Rev Cell Biol*, **2**:499-516.

Walter, P. and Johnson, A. E. 1994. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. *Annu Rev Cell Biol*, **10**:87-119.

Wang, C. Z. and Chi, C. W. 2004. *Conus* peptides--a rich pharmaceutical treasure. *Acta Biochim Biophys Sin (Shanghai)*, **36**(11):713-723.

Wang, M., Yang, J. and Chou, K. C. 2005. Using string kernel to predict signal peptide cleavage site based on subsite coupling model. *Amino Acids*, **28**(4):395-402.

Watson, M. E. 1984. Compilation of published signal sequences. *Nucleic Acids Res*, **12**(13):5145-5164.

Watts, C., Wickner, W. and Zimmermann, R. 1983. M13 procoat and a pre-immunoglobulin share processing specificity but use different membrane receptor mechanisms. *Proc Natl Acad Sci U S A*, **80**(10):2809-2813.

Wei, M. L. and Cresswell, P. 1992. HLA-A2 molecules in an antigen-processing mutant cell contain signal sequence-derived peptides. *Nature*, **356**(6368):443-446.

Weihofen, A., Lemberg, M. K., Ploegh, H. L., Bogyo, M. and Martoglio, B. 2000. Release of signal peptide fragments into the cytosol requires cleavage in the transmembrane region by a protease activity that is specifically blocked by a novel cysteine protease inhibitor. *J Biol Chem*, **275**(40):30951-30956.

Weiss, J. B. and Bassford, P. J., Jr. 1990. The folding properties of the *Escherichia coli* maltose-binding protein influence its interaction with *SecB* in vitro. *J Bacteriol*, **172**(6):3023-3029.

Weltman, J. K., Skowron, G. and Loriot, G. B. 2007. Influenza A H5N1 hemagglutinin cleavable signal sequence substitutions. *Biochem Biophys Res Commun*, **352**(1):177-180.

Westers, L., Westers, H. and Quax, W. J. 2004. *Bacillus subtilis* as cell factory for pharmaceutical proteins: a biotechnological approach to optimize the host organism. *Biochim Biophys Acta*, **1694**(1-3):299-310.

Wickner, W. 1979. The assembly of proteins into biological membranes: The membrane trigger hypothesis. *Annu Rev Biochem*, **48**:23-45.

Wickner, W. 1980. Assembly of proteins into membranes. *Science*, **210**(4472):861-868.

Wiedmann, M., Kurzchalia, T. V., Bielka, H. and Rapoport, T. A. 1987. Direct probing of the interaction between the signal sequence of nascent preprolactin and the signal recognition particle by specific cross-linking. *J Cell Biol*, **104**(2):201-208.

Wiley, H. S. and Michaels, G. S. 2004. Should software hold data hostage? *Nat Biotechnol*, **22**(8):1037-1038.

Williams, E. J., Pal, C. and Hurst, L. D. 2000. The molecular evolution of signal peptides. *Gene*, **253**(2):313-322.

Wolfe, P. B., Zwizinski, C. and Wickner, W. 1983. Purification and characterization of leader peptidase from *Escherichia coli*. *Methods Enzymol*, **97**:40-46.

Wollenberg, M. S. and Simon, S. M. 2004. Signal sequence cleavage of peptidyl-tRNA prior to release from the ribosome and translocon. *J Biol Chem*, **279**(24):24919-24922.

Wu, C. M. and Chung, T. C. 2006. Green fluorescent protein is a reliable reporter for screening signal peptides functional in *Lactobacillus reuteri*. *J Microbiol Methods*, **67**(1):181-186.

Xue, H., Lu, B. and Lai, M. 2008. The cancer secretome: a reservoir of biomarkers. *J Transl Med*, **6**:52.

Yamamoto, Y., Taniyama, Y., Kikuchi, M. and Ikehara, M. 1987. Engineering of the hydrophobic segment of the signal sequence for efficient secretion of human lysozyme by *Saccharomyces cerevisiae*. *Biochem Biophys Res Commun*, **149**(2):431-436.

Yamamoto, Y., Taniyama, Y. and Kikuchi, M. 1989. Important role of the proline residue in the signal sequence that directs the secretion of human lysozyme in *Saccharomyces cerevisiae*. *Biochemistry*, **28**(6):2728-2732.

Ye, R. D., Wun, T. C. and Sadler, J. E. 1988. Mammalian protein secretion without signal peptide removal. Biosynthesis of plasminogen activator inhibitor-2 in U-937 cells. *J Biol Chem*, **263**(10):4869-4875.

York J, Romanowski V, Lu M, Nunberg JH. 2004. The signal peptide of the Junín arenavirus envelope glycoprotein is myristoylated and forms an essential subunit of the mature G1-G2 complex. *J Virol*, **78**(19):10783-92.

Zhang, Z. and Wood. W. I. 2003. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics App Notes*, **19**(2):307-308.

Zhang, Z. and Henzel, W. J. 2004. Signal peptide prediction based on analysis of experimentally verified cleavage sites. *Protein Sci*, **13**(10):2819-2824.

Zheng, N. and Gierasch, L. M. 1996. Signal sequences: the same yet different. *Cell*, **86**(6):849-852.

Zheng, R. Y. 2004. Biological applications of support vector machines. *Brief Bioinform*, **5**(4):328-338.

Appendix A: Standard Amino Acid Abbreviations

Name of Amino Acid	3-Letter Code	1-Letter Code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Appendix B: SP Filtering Rules (Version 2.0)

The collection of rules listed here is a combination of several good practices proposed in previous works (Nielsen, *et al.*, 1996; Nielsen and Krogh, 1998; Emanuelsson, *et al.*, 2000; Menne, *et al.*, 2000; Chou and Shen, 2007; Plewczynski, *et al.*, 2008) and also newly formulated rules proposed along the course of this work. Applying this set of rules to the databases (see [A]) enables the generation of a preliminary filtered set of SPs with significantly reduced errors. The resulting filtered set will still require manual curation since there may be entries with inconsistency in annotation (e.g. an entry may not be tagged as containing putative results even if that is the case).

[A] Databases required:

(i) UniProt-KB/Swiss-Prot (exclude TrEMBL)

Organisms with these keywords are classified as **Gram-positive** bacteria:
Firmicutes, Actinobacteria, Deinococcus-Thermus, Fibrobacteres, Thermotogae, Chloroflexi, Dictyoglomi

Organisms with these keywords are classified as ***Gram-negative*** bacteria:
Proteobacteria, Planctomycetes, Fusobacteria, Acidobacteria, Chlorobi, Spirochaetes, Bacteroidetes, Cyanobacteria, Aquificae, Chlamydiae, Verrucomicrobia

(ii) EMBL

EMBL data categories:
(http://www.ebi.ac.uk/embl/Documentation/Release_notes/current/relnotes.html)

Entries belonging to these data groups are retained for integration:
Fungi, human, invertebrate, mouse, organelle, plant, prokaryote, rodent, viral, mammals and vertebrate

Entries belonging to the data groups are omitted:
Expressed sequence tags, bacteriophage, genome survey sequences, high-throughput genome sequences, unfinished DNA sequences generated by high-throughput sequencing, patent sequences, synthetic sequences, contig sequences and unclassified.

(iii) Protein Data Bank (PDB)

[B] Detailed procedures:

1. Retain only entries tagged with the SIGNAL keyword in the feature table **FT** field (http://www.expasy.org/sprot/userman.html#FT_line). This essentially omits mTP and cTP since transit peptides are identified by the keyword TRANSIT

2. Entries that are found

   WITHOUT

   - Accession number (**AC**)
   - date of creation or last annotation (**DT**)
   - taxonomic classification (**OC**)
   - SIGNAL keyword (**FT**)
   - sequence data (**SQ**)
   - Met as the starting residue (**SQ**)
   - Mature peptide portion (**SQ**)

   or

   WITH

   - fragment (**DE**)
   - organellar proteins (**OG**)
   - cell wall e.g. mollicutes (**OC**)
   - PROKAR_LIPOPROTEIN (**DR**) – they are cleaved by SPase II-cleaved lipoprotein SPs (Taylor, *et al.*, 2006)
   - Tat-type signal (**FT**) – rely on different mechanism for processing cleavage site (Blaudeck, *et al.*, 2001)
   - not cleaved (**FT**)
   - non-standard amino acids as identified by the characters 'X', 'Z' or 'U' found in sequence

   are all omitted from further parsing

3. Entries annotated with keywords such as PROBABLE, POTENTIAL, BY SIMILARITY, HYPOTHETICAL, MISSING, INFERRED, PUTATIVE AND CONFLICT are tagged to be *unverified*

4. Entries with ambiguous positions (either at the cleavage site or at the starting position) are designated as *unverified*. Such entries may be due to its sequence being partially sequenced. It may also be the case where some of these positions were not determined in the experiment. Part of the MP region that is used in the entry is also checked for such ambiguity.

5. SPs with length less than 11aa are tagged as *unverified* set since SPs are generally considered to be of length 15 to 40 with the shortest being 11aa

6. Use the 1st cross-reference under EMBL field in Swiss-Prot entry to automatically integrate the information from EMBL database. Those entries without any EMBL reference are removed.

   Swiss-Prot entries with *status identifiers* that appear in the **DR** field are sent for manual curation (http://www.expasy.org/sprot/userman.html#DR_line):

   (i) lack of annotations in the EMBL entries;

   (ii) indicated with annotation such as NOT_ANNOTATED_CDS, ALT_INIT, ALT_SEQ in their EMBL cross-references

7. The fields from EMBL: *sig_region* and *misc* are checked against the Swiss-Prot entries. This enables identification of inconsistency in positions quoted by either sources