Finding Secretion Signals in Protein Sequences

MASTER'S THESIS

Emmanuel Ojefua
Institute of Medical Technology
University of Tampere

August 2009

Preface

This research work was written in the Institute of Medical Technology, University of Tampere, Finland, 2008 – 2009. The practical work was done in the Bioinformatics/ Structural Genomics Laboratory in the Institute of Biotechnology, University of Helsinki.

I would like to thank my supervisors, Prof. Liisa Holm and Marrti Tolvanen for their guidance and support throughout the various stages of this master's thesis work. My appreciation also goes to Prof. Mauno Vihinen of the Institute of Medical Technology, University of Tampere, Matti Kankainen, and other members of Holm's group in the University of Helsinki. My colleagues and friends in the University of Tampere are also appreciated here for their useful suggestions and support. Finally, I express my sincere gratitude to my family for their love and warm support.

Emmanuel Ojefua

MASTER'S THESIS

Place:	UNIVERSITY OF TAMPERE
	Faculty of Medicine
	Institute of Medical Technology
Author:	OJEFUA, EMMANUEL EROMOSELE

Title:	Finding Secretion Signals in Protein Sequences
Pages:	77pp
Supervisors:	Prof. Liisa Holm and Martti Tolvanen, Ph.Lic.
Reviewers:	Prof. Liisa Holm and Prof. Mauno Vihinen
Date:	August 2009

Summary

Background and aims: Signal peptides are central to biological processes in that they direct proteins to their proper destination after synthesis. If the signal sequence in a nascent was changed, the protein could end up in a wrong cellular location. Some proteins known to be secreted have had their signal peptides unpredicted by existing signal peptide prediction tools. The aim of this study is to pre-process these proteins in order to optimize their start sites and apply a combined set of tools to finding their signal peptides.

Methods: The methods used here took a step backwards by starting at the DNA level of a given protein. DNA flanks were extended to allow for alternative gene prediction before translating back to proteins. The optimized protein sequences were analyzed using a set of signal peptide analysis tools. The results of these were compared with from the original protein sequences.

Results: Signal peptide prediction was enhanced in the optimized sequences. An actual prediction of the presence of signal peptide was seen in 2 out of the 5 proteins studied.

Conclusion: The results indicate that the methodology used in this study can help in future design and analysis of signal peptides. It also shows that the hypothesis that non-classical proteins could be predicted by using alternative gene predictions is supported.

List of Abbreviations

ANN	Artificial neural network
EF-Tu	Elongation factor Tu
EST	Expressed sequence tag
GroEL	Growth elongation factor
GTP	Guanine triphosphate
HHM	Hidden Markov Model
IMP	Integral inner membrane protein
MSA	Multiple sequence alignment
NN	Neural Network
ORF	Open reading frame
RBS	Ribosomal binding site
SCL	Subcellular localization
SOM	Self organizing map
SP I	Signal peptidase I
SP	Signal peptide
SRP	Signal recognition particle
SVM	Support Vector Machine
TAT	Twin-arginine protein transport
TM	Transmembrane

Contents

1 Introduction

The availability of genome sequences of whole organisms have placed us in a position to understand the expression, function, and regulation of the entire set of proteins encoded by an organism. This information is invaluable for understanding how complex biological processes occur at a molecular level, how they differ in various cell types, and how they are altered in disease states.

Bacterial cells generally consist of a single intracellular compartment surrounded by a plasma membrane. In contrast, eukaryotic cells are elaborately subdivided into functionally distinct, membrane bound compartments. Most eukaryotic proteins are encoded in the nuclear genome and synthesized in the cytosol, and many need to be further sorted before they reach their final destination. The localization of a protein is largely determined by a trafficking system that is reasonably well understood for some organelles in which there are two main branches (Nair and Rost, 2008).

Proteins are synthesized, on one branch, on cytoplasmic ribosomes, and can proceed from there to the nucleus, mitochondria, or peroxisomes. The second branch leads from the endoplasmic reticulum-ribosomes to the Golgi apparatus and from there to lysosomes, or secretory vesicles, and to extracellular space. Many proteins destined for the branch point leading to secretion have an N-terminal signal peptide which is cleaved off proteolytically either during or after protein translocation through the membrane. Proteins lacking this signal are retained in the cytoplasm. The targeting signals used at the other branch points are not always so clear for two reasons. First, the signals used are presented by folded proteins, and hence are not always contiguous in sequence. Second, even where the sequences are contiguous, not all signal peptides have been documented (Nair and Rost, 2008).

Signal peptides direct proteins to their proper cellular and extracellular locations. One major example of such a process is the translocation of proteins across the well-established sec pathway found in both eukaryotes and prokaryotes where proteins designated for export from the cell are tagged by an N-terminal signal sequence. This

signal sequence directs the protein across the cell membrane and it is usually cleaved off by a signal peptidase after translocation. Signal peptides for the Sec pathway generally consist of a positively charged N-region, a hydrophobic h-region, and an uncharged but polar C-region. While the cleavage site for the signal peptidase is located in the c-region, the degree of signal sequence conservation, as well as the cleavage site position, however, varies significantly between different proteins (Hiller *et al,* 2004).

In the absence of a clear understanding of the principles governing protein translocation, computational methods for predicting subcellular localization have pursued a number of conceptually distinct approaches. Methods for predicting protein sorting signals have primarily explored the following avenues; annotation transfer from homologous sequences, predicting the sorting signals that the cell uses as address labels, mining the functional information deposited in databases and scientific literature, and using the observation that the subcellular localization depends in subtle ways on the amino acid composition of the protein. There are also methods that combine the outputs from a number of primary methods to optimize accuracy and coverage.

Since protein trafficking relies on the presence of sorting signals, it would be naturally ideal to predict the signal responsible for targeting. However, the difficulty in accurately identifying sorting signals is largely due to the complexity of the mechanism(s) involved in protein sorting in nature. Despite their limited applicability, methods that predict sorting signals provide the most useful predictions since by pinpointing the targeting signal they shed light on the molecular mechanism of protein translocation.

1.1 Goals of the thesis

This Master's thesis presents a method for finding secretory signals in protein sequences. Some proteins are known to be secreted yet existing prediction tools have not been able to predict any signal peptides from them. Part of the problem with wrong predictions is also associated with wrong start sites annotation of protein sequences in databases where data is derived for benchmarking most of the existing

tools. Here, an attempt is made to optimize these start sites by screening proteins resulting from alternate gene prediction.

First, a review of literature is presented in the next chapter. In chapter three, the biological motivation for the study is discussed while the methodology and approach to the research is presented in chapter four. The results obtained from the analysis are presented in chapter 5 which are then discussed in chapter 6, and possible future developments and research goals highlighted. Based on the above, a conclusion is drawn in chapter 7.

2 Literature review

The targeting of proteins to their different cellular locations in the presence or absence of signal sorting peptides has been well studied and there is a vast amount of literature describing them. The increase in the volume of results obtained from experimental work has hugely necessitated the idea of developing novel bioinformatic methods for effective analysis and interpretation of these results.
This review attempts a brief description of the biological concepts of the subject while emphasis is laid on the architecture, performance, and comparisons of some of the existing bioinformatic tools related to secretory signal prediction.

2.1 The biological membrane

Biological membranes are thin structures that form the boundary between the exterior and interior of the cell, as well as between different cellular compartments. Biological membranes consist of a mixture of lipids and proteins. Membrane lipids assemble into bilayers which functions permeability barrier that prevents leakage and enables different chemical environments to exist on each side of the membrane. This is a fundamental aspect of life since many biological processes depend on the formation of concentration gradients across a membrane.

However, for a cell to survive, it is essential that membranes permit passage of various molecules such as nutrients and ions. Proteins are therefore essential components of the biological membrane since they can function as pores, channels or transporters that allow selective passage across a bilipid layer (Von Heijne, 1990).

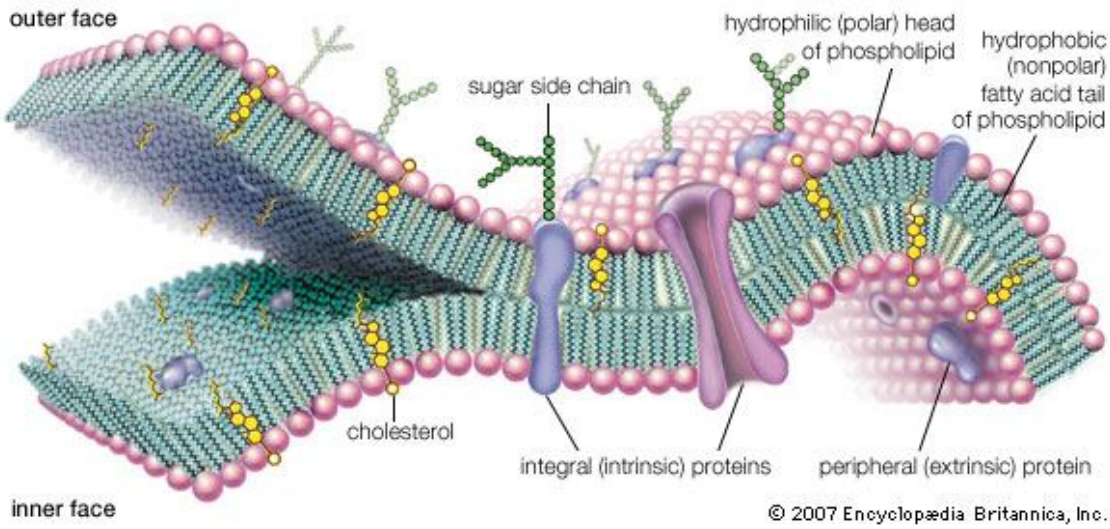

Figure 2.1 The biological membrane (Encyclopaedia Britannica, 2007).

2.2 Secretory and non-secretory proteins

Secreted proteins of bacteria are those released to the extracellular milieu, whereas exported proteins are localized in the periplasm or cell wall (Pugsley, 1993). The large majority of these proteins possess an N-terminal signal sequence that mediates their membrane translocation via the Sec-depedent translocation.

Given a protein sequence, the first step in predicting its signal peptide is to identify whether it is a secretory protein or non-secretory protein. For the latter, no prediction is needed at all because it contains no signal peptide (Liu *et al*, 2007). A large number of the existing methods are characterized by their ability to effectively classify proteins as secretory or non-secretory.

2.3 Signal peptides

A signal sequence is a short peptide that functions as an “address tag” in directing a nascent protein to wherever it is supposed to be. If the signal sequence in a nacent protein

was changed, the protein could end up in a wrong cellular location causing various diseases (Liu et al, 2007).

The best known protein "zip code" is the secretory signal peptide, which is found in all the three domains of life (Von Heijne, 1990). It targets a protein for translocation across the plasma membrane in prokaryotes and across the endoplasmic reticulum (ER) membrane in eukaryotes (Von Heijne, 1990).

2.3.1 Structure

Signal peptides typically have three distinct domains, including an amino terminal positively charged region (N-region, 1-5 residues long); a central, hydrophobic part (h-region, 7-15 residues); and a more polar carboxy-terminal domain (C-region, 3-7 residues). Beyond this overall pattern, no precise sequence conservation could be found, and it soon became obvious that signal peptides are highly variable, rapidly evolving structures (von Heijne, 1990). Variations in mean lengths and amino acid compositions of the three regions between different groups of organisms have been detected (von Heijne and Abrahmsen, 1989). The n-, h-, and c- regions of signal peptides from eukaryotes tend to be slightly shorter than those from Gram-negative bacteria. Due to the composition and chemical nature of the membrane, it has been suggested that the positive-hyrophobic-polar design of signal peptides might efficiently bind to membrane lipid bilayers in a loop-like structure with basic amino acid terminus aggregating to acidic lipid head groups on the membrane's cytoplasmic face (Engelman and Steitz, 1981). But there is still no consensus whether signal peptides interact primarily with receptor proteins or directly with the lipid bilayer itself (von Heijne, 1990).

2.3.2 Protein secretory pathways

Signal sequences can direct proteins to a membrane through different targeting pathways and select different translocation systems for the actual transport across the memrane.

2.3.2.1 Sec pathway

The Sec B and the signal recognition particle targeting pathways converge at the Sec-translocon which functions as a protein conducting channel in the inner membrane. The Sec-translocon is generally assumed to be required for translocation of most secretory proteins across, and insertion of most integral inner membrane proteins (IMPs) into, the inner membrane of *Eschericia coli*.

Targeting and translocation can occur co- or post-translationally and can either be dependent on the signal recognition particle (SRP) and docking protein/SRP receptor or be SRP independent; in the latter case, targeting and translocation involve the SecB protein in bacteria and Sec62p-Sec63p complex in yeast (Walter and Johnson, 1994). The two targeting pathways converge at the Sec translocon at the membrane, assembled from the SecA and the SecY-E-G in bacteria. The selection of the SRP-dependent or SRP-independent pathway is determined largely by features of the signal sequence. The h-region of a signal sequence has been implicated as the parameter discriminating between SRP-dependent and SRP-independent pathways according to studies in yeast (Walter and Johnson, 1994). Signal sequences directing proteins into the SRP-mediated pathways have a significantly more hydrophorbic h-region than those mediating SRP-independent targeting. Discrimination between the two targeting pathways in bacteria seems to have the same features as above (Martgolio and Dobberstein, 1998).

2.3.2.2 TAT pathway

A signal-sequence-mediated transport system that is independent of the Sec apparatus has been identified in the thylakoid membrane of chloroplasts. Proteins are exported via the Sec-pathway in an unfolded state. In contrast, the Twin-Arginine protein transport (TAT)-pathway is used for post –translational translocation of folded proteins across the inner membrane of *E.coli* (Lee *et al*., 2006). The TAT-translocon in E.coli is composed of the integral inner membrane proteins TatA, TatB, and TatC proteins whose homologs exist in Archaea, prokaryotes and

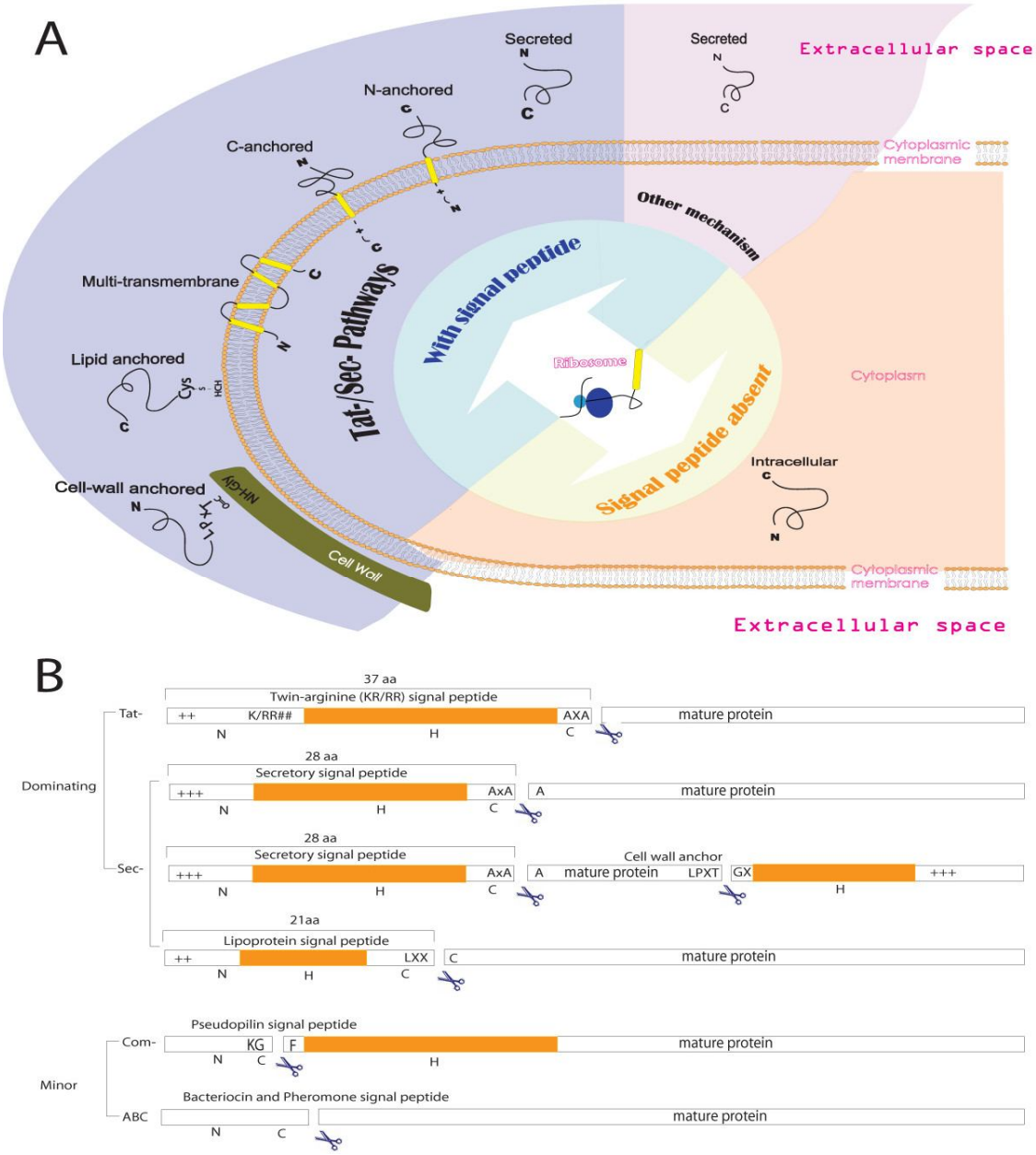

Figure 2.2 (A) Classification of protein subcellular localization in Gram-positive bacteria (B) The structure of known signal peptides. Tat- and Sec-dependent signal peptides have their overall structure commonly conserved as distinctive N, H and C regions (Adapted from Zhou *et al*., 2008).

plant mitochondria. Results from biochemical studies suggest that TatA, TatB, and TatC participate in dynamic, multimeric membrane complexes that interact transiently

witheach other and the substrate during the translocation cycle (Sargent *et al*., 2006). The initial recognition and binding of TAT signal peptide is thought to be mediated by the TatBC complex while the formation of the protein-conducting channel is believed to be due to the oligomeric TatA. Some evidence suggesting that the protein translocation step is also dependent on proton motive force is available (Sargent *et al*., 2006). This proton motive force is derived from the induction of an electrochemical gradient as a result of both the electric potential and chemical potential across a membrane.

Substrates are directed to the TAT-translocon by a distinct twin-arginine motif present in the signal sequence. A subset of TAT signal peptides exhibits a high degree of pathway selectivity, while others are promiscuous and can direct proteins either to the Sec- or TAT- translocon (Lee *et al*., 2006). A number of TAT-dependent proteins that do not have a signal sequence have been isolated, and they form complexes with TAT-signal-containing proteins . These proteins eventually get secreted through the TAT-pathway by a “hitch-hiker mechanism” (Lee *et al*., 2006). IMPs having a single TM domain at their carboxy-terminal end seem to depend on the TAT-translocon for insertion. The reason for this might be due to the fact that the TAT-translocon a common “Ala-X-Ala” motif at the C-region. This region usually precedes the cleavage site (Tuteja, 2005).

2.3.3 Membrane anchors

Signal sequences can anchor proteins in the membrane when they contain a sufficiently hydrophobic h-region and are not cleaved by signal peptidase. They are then called ‘signal-anchor’ sequences to indicate their dual role in targeting and membrane anchoring (High and Dobberstein, 1992). Signal anchor sequences can insert into the membrane in orientation, keeping the N-terminus on the cytoplasmic side or transferring it across the membrane (High and Dobberstein, 1992). During biosynthesis of a type II membrane protein, the nascent polypeptide with the signal-anchor sequence forms a loop across the membrane, and the C-terminus of the protein is transferred across the membrane while the N-terminus remains on the cytoplasmic side. In contrast, the signal-anchor sequences

of type I membrane protein transfers the N-terminus across the membrane while the C-terminus remains in the cytoplasm (Martoglio and Dobberstein, 1994).

2.3.4 Signal sequence cleavage

The signal peptidases are unique serine proteases that carry out catalysis using a serine/lysine dyad instead of the prototypical serine/histidine/aspartic acid triad found in most serine proteases (Tuteja, 2005). They are responsible for the cleavage of signal peptide of many secreted proteins in bacteria and serve as a potential target for the development of novel antibacterial agents due to their unique physiological and biochemical properties. The catalytic domain of signal peptidases (SPase) extends into the periplasmic space and is anchored to the membrane by two membrane segments located at the N-terminal end of the protein (Tuteja, 2005).

The substarte specificity of the SPase is similar in prokaryotes and eukaryotes suggesting that both the signal peptides and SPase specificity are conserved throughout evolution. Preproteins contain a signal sequence with a positively charged amino-terminus, a central hydrophobic domain, and a neutral but polar but polar C-terminal domain. Signal peptide cleavage site is specified by the c-region while the n- and h- regions are required for translocation. The signal-peptide cleavage site specificity is often designated “Ala-X-Ala” rule due to the presence of of Ala at the -3 and -1 position. Despite having no distinct consensus sequence other than a commonly found c-region “Ala-X-Ala” motif preceeding the cleavage site, signal sequences are recognized by SPase I with higher fidelity. Alternatively, the eukaryotic signal peptides show a less stringent criteria of the availability of Ala at these positions, and thus can allow Gly, Ser, Thr, and Cys with a frequency similar to Ala at the -1 position and Ile, Leu, Val, Ser,, and Thr with a frequency similar to Ala at position -3 (Tuteja, 2005).

The signal sequence signals the cellular fate or destination of a newly synthesized protein directing it to its ultimate destination in the cell.

2.3.5 Significance of signal peptides

The significance of signal peptides and their possible applications in biotechnology are highlighted below.

2.3.5.1 Drug discovery and therapeutics

Knowledge of the subcellular localization of a protein can significantly improve target identification during the drug discovery process. Drug molecules easily access secreted proteins and plasma membrane proteins since they are localized in the extracelluar space or the cell surface. Also, secreted proteins or receptor extracellular domains that have been purified can be utilized directly as a therapeutic, or may be targeted by specific antibodies or small molecule. Therapeutic targeting proteins present on the cell surface in a specific cell type or disease state has been created. For example, Rixtuan is an antibody therapeutic targeting the B lymphocyte-specific CD20 protein and, is effective in the treatment of non-Hodgkin's lymphoma. Aberrant subcellular localization of proteins has been observed in the cells of several diseases such as cancer and Azheimer's disease (Nair and Rost, 2008).

2.3.5.2 Elucidation of biological processes

A signal sequence functions as an "address tag" in directing protein to wherever it is supposed to be. If the signal sequence in a nascent protein was changed, the protein could end up in a wrong cellular location causing various diseases. Therefore knowledge of signal sequences can be used to reprogram cells in a desired way for future cell and gene therapy (Liu *et al*., 2007).

Secreted proteins and integral plasma membrane proteins are of special interest since they play key roles in important biological processes, e.g., signal transduction and transmission, and cellular differentiation. A knowledge of which proteins are native to the cytosol and those that are targeted to an organelle is extremely useful in assembling

metabolic pathways that putatively occur in the organelles in question (Schneider and Fechner, 2004).

2.4 Protein secretory signal prediction

Prediction methods for protein sorting signals can be roughly divided into two broad categories. We have those methods that use the amino acid sequence of the protein exclusively as the input and those that also require some other additional input data, for instance, data obtained from expression level, phylogenetic profiles, and database entries with respect to their lexical context as well as Gene ontology terms (Chou and Shen, 2006).

Over time, performance has been reported to be better when additional information is provided when compared with just sequence-based methods, whereas the applicability will be more limited due to the fact that these methods can only be used for examples where this additional information is available (Chou and Shen, 2006). It is therefore important that the use of lexical context or GO-terms requires that the proteins are already to some extent annotated, and this annotation may often include information relating to subcellular localization, either explicitly or implicitly.

GO-numbers utilizing methods are technically able to accept input consisting of sequences only, but it should be noted that their high reported performance has been measured only on data sets where almost all sequences had GO-numbers and a high proportion of these were of the "cellular component" type (Chou and Shen, 2006).

There are also methods requiring only a sequence as input which is used to search databases for homologues or to look for ocurrence of certain protein domains (Scott *et al*., 2004). Annotation by homology to proteins with known subcellular location can yield a good prediction comparable to that of machine learning methods.

Extensive analysis has been carried out by Nair and Rost, (2008) on how close two proteins should be in sequence space to have the same subcellular localization. If we infer homology from pairwise identity, the conclusion was that over 70% sequence identity is needed to correctly infer localization for 90% of all proteins. In this regard, the BLAST expectation value or the so-called HSSP distance was found to be better than percent identities for localization predictive accuracy. With respect to E-value, its natural logarithm should ideally be below -80 to achieve a 90% level of accuracy.
Meanwhile, Yu et al (2006) have reported that localization prediction by homology was better than a machine learning method above a cut off.

2.4.1 Prediction of actual sorting signals

One of the first attempts at predicting subcellular localization was a weight matrix for secretory signal peptides (von Heijne, 1986). A weight matrix consists of a simple sequence profile generated from a multiple sequence alignment without gaps, where the amino acid counts at each position in a window around the area of interest readily provides the basis for the calculation of the weights. This has been found to be extremely useful and in wide usage too. The weight matrix is included in PSORT, but however does not exist as stand alone while it is still used in the SPscan and Sigcleave tools respectively.
HMMs of the well-known profile architecture can also be seen as an extension of weight matrices, where the alignment used to calculate the weights can contain gaps, so that motifs of varying length can be represented. In addition to the profile architecture, there are many ways to build an HMM; for example, a branched model can represent a choice between alternative patterns, while a cyclic model represents a repeating pattern. SVMs treat each input pattern as a set of numbers which is mapped onto a high-dimensional space by the so-called kernel function, and then define an optimal separating hyperplane in that space which classifies the patterns into two categories and maximizes their distance from the hyperplane (Byvatov and Schneider, 2003).
In SignalP's version 3, for example, the input to the neural network part has been extended by including, in addition to the moving amino-acid windows themselves, the

relative positions of each window and the overall amino-acid composition of the entire sequence (Byvatov and Schneider, 2003).

2.4.2 Prediction by global sequence properties

Prediction of protein sorting signals can also take advantage of the fact that proteins of different subcellular compartments differ in global properties which reflects in their amino acid make-up.
Although the signal prediction methods are probably closer to mimicking the actual information processing in the cell, methods based on global properties can work also for genomic or EST sequences where the N-terminus of the protein has not been included or correctly predicted. Another advantage is that they provide the opportunity to predict localizations for which the sorting signals are not known or not adequately defined. One drawback is that such methods will not be able to distinguish between very closely related proteins or isoforms that differ in the presence or absence of a sorting signal (Emmanuelsson *et al*, 2007).

The prediction of protein sorting by global sequence properties by using some odd-ratio statistics to discriminate between soluble intracellular and extracellular protein based on amino acid composition by using residue frequency was pioneered by Nakashima and Nishikawa, (1994). The NNPSL method by Reinhardt and Hubbard, (1998), pioneered the use of neural networks (NNs) trained on overall amino acid composition to predict localization. This method was able to distinguish three bacterial compartments, but performed poorly in predicting plant proteins. A number of methods that are built on SVM which utilize amino acid composition in predicting sorting signals also exist SubLoc being the first (Gardy and Brinkman, 2006).

Using amino acid composition as the only input will lead to discarding all information regarding the sequence order. Parts of this order can be incorporated into global property methods, for example by using the frequencies of amino acid pairs in a consecutive /dipeptide manner or separated by a number of positions.

PSORT is a widely used predictor of sub cellular localization utilizing both sorting signals and global features. A decision tree is used in the original version of PSORT to make predictions which involve computations and comparisons of a set of sequence derived parameters operating under some "localization rules" accruing largely from literature. Many of these rules relate to the presence of various sequence motifs that enable proteins to be localized to a certain compartment while others deal with amino-acid content in certain regions. PSORT discriminates between 17 different compartments for plants, 14 for animals and 13 for yeast.

2.5 Machine learning methods for signal prediction

Hidden markov Models (HMMs), supervised multilayer feed-forward artificial neural networks (ANNs), self-organizing maps (SOM), and support vector machines (SVMs) have been widely used for constructing rules that can be employed in determining parts of a protein sequence responsible for its targeting – the targeting signals (Byvatov and Schneider, 2003). Apart from HMMs, which can be built with only a set of "positive examples", like already known signal peptides, the above prediction systems usually represent nonlinear classifiers that separate "positive examples" from "negative examples".

Training and testing phases are two basic steps common to developing any machine learning method. In the training phase, classifiers are established by simply adapting internal model parameters while in the test phase, there is an assessment of the performance and generalization ability using defined statistical methods like jack-knifing or bootstrapping with data sets not used for training (Duda, *et al*., 2001).

2.5.1 Hidden Markov Models

Hidden Markov Models, neural networks, stochastic grammars, and Bayesian networks are closely related. The success of the HMM approach is critically influenced by an appropriate alignment of the training sequences (Scheidner and Fechner, 2004). A

standard HMM consists of a finite set of nodes representing “hidden states”. These nodes are interconnected by links describing the probabilities of a transition between the individual states. Additionally, each hidden state has an associated set of probabilities of emitting a particular “visible state”. A discrete alphabet A of symbols is assigned to the hidden and visible states. In the context of proteins sequences, A is the standard 20–letter amino acid alphabet. The transition matrix T specifies the probabilities of going from the hidden state x to the hidden state y. The emission matrix E indicates the probabilities of emitting a certain symbol S in a certain hidden state (Scheidner and Fechner, 2004).

The model parameters T and E are determined from a collection of training samples during HMM training and as such no learning method can guarantee achievement of optimal system parameters. However, algorithms exist that have been shown to be well suited for HMM optimization. Because HMMs allow for the modeling of sequences of varying lengths, it possesses an inherent property that is very useful in predicting signal sequences.

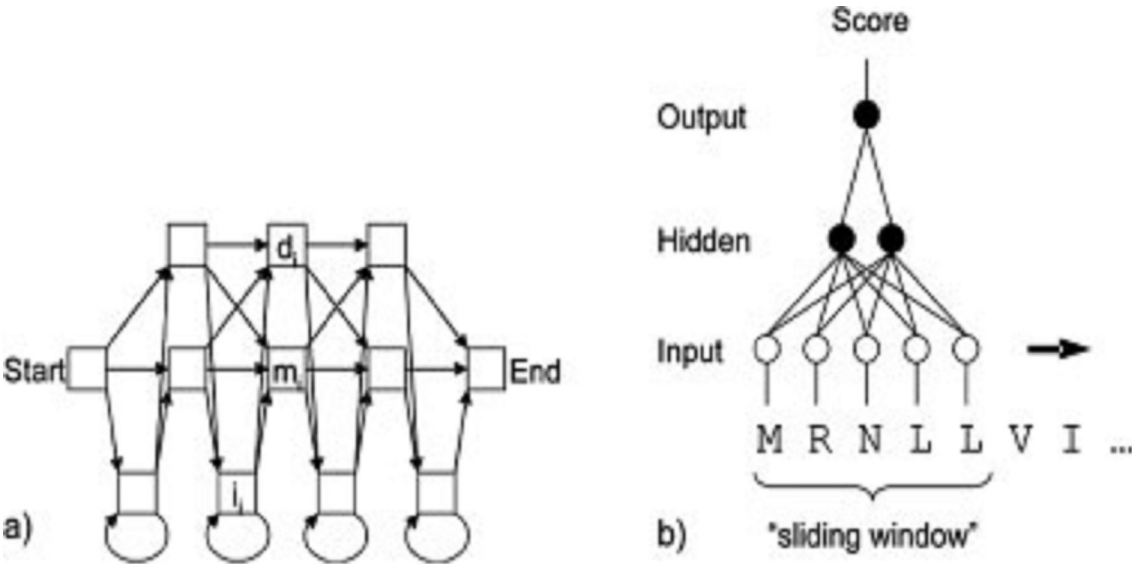

Figure 2.3 (a) Schematic of a standard HMM with *m* as the main state, *d* is deletion state and *i* is insert state. (b) Schematic of a three-layered ANN (Scheidner and Fechner, 2004).

In HMM, the number of parameters used for the model increases very quickly with respect to the size of the alphabet employed. This number can be unfavourably large in

the case of a huge amount of protein models and this is a setback for HMM. Another drawback is their inability to express dependencies between non-neighboured hidden states. This can prevent the identification of complex signal sequence features that might be formed by interaction of non-neighboured residues (Schneider and Fechner, 2004). A possible solution to this might be the use of correlation-based descriptors for sequence analysis and other modeling algorithms than HMM (Chou, 2000). Despite the above disadvantage, HMM models have achieved major success when applied to the prediction of protein targeting signals.

Krogh and coworkers have developed an HMM used for the prediction of lipoprotein signal peptides which are cleaved by SP II, which is a specific signal peptidase for precursors of bacterial lipoproteins. They found the prediction accuracy to be greater than 98% with only 0.3% false positive assignments of other targeting signal-containing sequences when assessed by a leave-one-out statistics with 63 proteins (Juncker *et al*., 2003). Another HMM model which yielded approximately 95% sensitivity and specificity for eukaryotic signal peptide prediction using a collection of 892 human and 644 mouse-signal-containing proteins has been developed by Zang and Wood, (2003).

2.5.2 Self-organizing map

SOMs or Kohonen- networks are unsupervised neural networks (Kohonen, 1982). SOMs try to map in a non-linear manner a high-dimensional input space to a lower-dimensional input space resulting in the proximity of points in the target space reflecting proximity of points in the source space.

They consist of an input and an output layer where the number of input neurons and the number of weights connected to an output neuron is identitical to the dimension of the input data. In SOMs, data clusters are represented by each output neurons where the number of output neurons amounts to the number of clusters. Every data point is mapped to exactly one output neuron and this mapping is determined by the best match of an existing data point, for example, a protein sequence with a weight vector of an output

neuron. These weight vectors in SOMs are actually determined at the training phase (Kohonen, 1982).

SOMs have been utilized in comparative proteome analysis as well as secondary structure prediction (Kohonen, 1982). The use of SOMs, as a public web based prediction tool for finding protein targeting signals does not yet exist even though several methods utilizing SOM to predict signal peptides have been developed (Schneider *et al*., 2009). However, SOMs have been used favourably to cluster and visualize proteins which are targeted to the extra cellular matrix and mitochondrion (Schneider, 1999).

2.5.3 Multilayer feed-forward networks

In supervised neural networks, the data used for training the network has to include information regarding the property the trained neural network should predict. ANNs are good examples of supervised neural networks.

ANNs are made up of formal neurons and interconnections between the former which are arranged in layers where at least 3 layers are needed to satisfy the formation of a multilayer feed-forward network. The input layer is usually the first layer. The layers in between the input layer and last are referred to as “hidden” layers, where the number of neurons can be adjusted based on the type of classification task. The number of dimensions of the input data is equal to the number of neurons in the input data (Schneider, 1999).

Every neuron of one layer is connected to every neuron of the next layer in a manner that no connections between the neurons of the same layer exist. These connections between neurons are numerical weight values that are optimized during network training according to an error function which tries to explain deviations of predicted target values from the observed values.

When properly trained, ANNs offer several advantages. They are, for example, able to cope with noisy data, and they have the ability to generalize (Schneider and Wrede,

1998). They are very qualified for modeling nonlinear input/output relationships. Due to the fact that understanding the decisive features of a trained neural network can be often problematic, ANNs are considered weak in this regard (Sadowski, 1998). There has however been the use of a different kind of concept to solve the problem of understanding and interpreting the rules and features used in training a given neural network by relying on probabilistic motif generation (Gonnet and Lisacek, 2002). These hmotifs describe preferred patterns of amino acid residues and properties that can be used for database searching and sequence classification. These are approaches that are complementary in nature and can help in predicting targeting signals.

ANNs do also have a problem of overfitting (Schneider and So, 2003). Generally, a neural network with perfect prediction can be found for every classification problem. But such a network is not able to generalize anymore: it predicts the training data perfectly but it fails to classify data not used during training. In general, multilayer feed-forward neural networks have found widespread use for classification tasks. This observation also hold for the classification of targeting signals as most of the prediction tools available employ the ANN machine learning method (Schneider and Fechner, 2004).

2.5.4 Support Vector Machine

The classical SVM method is fundamentally driven by data with its resultant use in binary classification. They are generated basically by a two-step procedure which includes first mapping the sample data vectors to a high-dimensional space. The dimension of this space s significantly larger than the dimension of the original space of the data. The largest margin separating the data in the above hyperplane is then implemented by the algorithm. It has been shown that classification accuracy usually depends only weakly on the specific projection, provided that the the target space is sufficiently high-dimensional (Cortes and Vapnik, 1995). When we do have a case of difficulty in finding a separating hyperplane even in very high-dimensional space, a tradeoff is introduced between the size of the separating margin and penalties for every data vector which lies within the margin (Cortes and Vapnik, 1995).

Points classified by SVMs can be divided into two groups, support vectors and nonsupport vectors. Non support vectors are classified correctly by the hyperplane and are located outside the separating margin. Parameters of the hyperplane do not depend on them, and even if their position is changed the separating hyperplane and margin will remain unchanged, provided that these points will stay outside the margin. Other points are support vectors, and they are points which determine the exact position of the the hyperplane. Informally speaking, support vector contain the important information for the classification task (Schneider and Fechner, 2004).

A major advantage of support vector machines is that it tends to be less susceptible to overfitting when compared to other classification methods. SVMs try to provide a solution where the separating hyperplane does not depend on the complete data set but solely on the support vectors. Instead, it looks out for one out of the many possible separating hyperplanes and finds one with the largest margin.

With respect to noisy data and features, SVM also come in handy with a handful of robust features (Cristianini and Shawe-Taylor, 2000). However, even with a seemingly smaller chance of overfitting than ANNs, the problem is not entirely absent from SVMs. Also, SVMs are not yet as widely and frequently used as ANNs.

The PSORT-B prediction tool employs an SVM to discriminate between cytoplasmic and noncytoplasmic sequences (Gardy *et al*, 2003). A more comprehensive use of SVMs has been described by Park and Kanehisa, (2003) who considered 12 subcellular localizations in eukaryotes for a prediction system (PLOC) consisting of 60 different SVMs and a jury decision. Sequence encoding was achieved by utilizing amino acid frequencies and residue-pair frequencies, yielding an overall accuracy of 80%. The study above produced outcomes which suggested that a good combination of different classifiers can result in more robust predictions when compared to individual systems.

The above study also complements that of earlier works by Chou and Elrod, 1999, who used covariant discriminant analysis methods to achieve classification but got slightly lower prediction accuracy.

2.6 Signal peptide prediction tools

Many software tools have been developed for *ab initio* cellular localization prediction, using machine learning techniques such as neural networks, hidden Markov models and support vector machines. Identifying the program best suited for a researcher's needs requires familiarity with several different programs. Prediction accuracy depends on the methods employed by a program and the integrity of the data used to develop the program (Klee and Ellis, 2005).

As we shall see in the following sections, several signal peptide prediction tools exist. These tools have been developed based on different techniques and approach. A major important consideration when developing tools is to use a technique that give rise to the best prediction as well as minimize error and or false positive predictions. On the other hand, the data used in the initial training of the tool enables one to set the tool’s parameters. Data sets for training and testing are as such very important to the outputs from the prediction tool. Tables 2.1 and 2.2 provides a summary of techniques and data sets.

Table 2.1 Summary of prediction tools showing technique used, prediction, and target organism.

Tool	**Technique used**	**Prediction**	**Applicable organism**	**Reference**
SignalP 3.0	HMM, NN	Cleavage site; signal peptide	Gram positive, Gram negative bacteria; eukaryotes	Bendtsen *et al*., 2004
Phobius	HMM	TM; Signal peptide	Prokaryote; eukaryotes	Bendtsen *et al*., 2004
WoLF PSORT	Multi-Component; k-Nearest neigh-bour	SCLs	Animal, plant, And fungi	Horton *et al*., 2007
PSORTI	Multi-component; k-nearest neigh-bour	SCLs	Gram positive and Gram negative Bacteria	Boeckman *et al*., 2003
PSORTb	SVM	SCLs	Gram positive and Gram negative bacteria	Gardy *et al*., 2006
PSORTII	k-nearest neighbour	SCLs	Bacteria, yeast, yeast/animal	
LocateP	Combination of numerous existing techniques	SCLs; pathways for Sec-or Tat-dependent secretion	Bacteria	Zhou *et al*., 2008
SigFind	NN	Signal peptides	Humans; other eukaryotes	Gardy *et al*., 2006
CELLO	SVM	SCLs	Gram positive and gram negative bacteria	Yu *et al*., 2006
SubLoc	SVM	SCLs	Prokaryotes; eukaryotes	Gardy *et al*., 2006
Proteome Analyst	Annotation keywords	SCLs	Gram positive and Gram negative bacteria	Gardy *et al*., 2006

Table 2.2 A summary of some tools showing training/testing data sets

Tool	**Dataset**	**Number of sequences (Training/Test)**	**Source of annotation**	**Reference**
SignalP 3.0	Menne; Dataset constructed by Menne	2708 Eukaryotic; 692 Gram positive bacterial 304 Gram negative bacterial	SwissProt	Bendtsen,*et al.*, 2004
Phobius	Menne	Positive test set: 426 Prokaryotic 1142 eukaryotic Negative test set: 398 Prokaryotic 1128 eukaryotic 105 Transmembrane	SwissProt	Menne et al, 2000
Psort		336 *E.coli* proteins (classified into 8 classes; accuracy, 81%); Yeast proteins (classified into 10 classes, accuracy, 55%)	SwissProt	
PsortB	ePsortDB	1443 proteins; cytoplasmic (248), inner membrane (268), perisplasmic(244), outermembrane(352), Extracellular(190), multiple sites(141)	Manually curated	Gardy *et al*., 2003
PsortII		1531 yeast sequences	SwissProt	Boeckman *et al*., 2003
WoLF PSORT		2113 Fungal; 2333 plant; 12771 animal	UniProt	Horton *et al*., 2007
LocateP		1370 Bacterial test sequences; 1336 positive prediction	SwissProt	Zhou et al., 2008
CELLO	Park & Kanehisa	1444 Gram negative bacterial (PS data)	SwissProt	Yu *et al*., 2006

Tool	Dataset	Number of sequences (Training/Test)	Source of annotation	Reference
		7589 Eukaryotic (Park and Kanehisa)		
SubLoc	Reinhardt & Hubbard	291 prokaryotic, 2427 eukaryotic	SwissProt	Lu *et al*., 2003
Proteome Analyst		16284 animal, 3420 plant 2104 fungi, 3218 Gram negative bacterial, 1571 Gram positive bacterial	SwissProt	Lu *et al*., 2003
TargetP		940 plant, 2738 non-plant	SwissProt	Lu *et al*., 2003
Sigfind	Menne	negative test 50 (2 FP returned) Positive test 50 (1 FP returned)	SwissProt	Bendtsen *et al*., 2004
LOCkey		1161 assorted(87% accuracy)	SwissProt	Lu *et al*., 2003

2.6.1 SignalP 3.0

The development of computational methods for the prediction of N-terminal signal peptides many years ago were initially based on utilizing a weight matrix approach (von Heijne, 1986). The SignalP tool predicts the presence of signal peptidase I cleavage sites. In lipoproteins, cleavage of peptide sequences from the transported protein involves the peptidase II enzyme. The presence of which has been predicted by the construction of the tool, LipoP (Juncker *et al*, 2003). While most methods only tries to classify proteins as

secretory or non-secretory, SignalP has been able to achieve prediction of both classification and assignment of cleavage site between the underlying amino acids.

SignalP since it was first developed has been periodically upgraded. The new version of SignalP, has been recently trained by generating a new, thoroughly curated dataset based on the extraction and redundancy method of Nielsen *et al*, (1996). By employing a couple of other methods to clean the new data set, Jannick *et al*, (2004), found a surprisingly high error rate in Swiss-Prot, where , for example, of the order of 7% of the Gram-positive entries had either wrong cleavage site position and/or wrong annotation of the experimental evidence.

2.6.1.1 Neural network architecture

Two different neural networks have been used used in the SignalP tool for coping with the signal peptide prediction problem. The cleavage site is recognized by one network and the other is used to determine whether a given amino acid belongs to the signal. Improvements to the neural networks were made by introducing new input features, such as the position of the sliding window as a parameter alongside the amino acid composition of the entire sequence. The signal peptide discrimination and the sigal peptidase cleavage site were handled using two different types of neural networks which was the same approach in the earlier signalP version (Nielsen et al, 2003)

The brute-force approach was used to optimize the window sizes for the neural networks by calculating single-position correlation coefficients for all possible combinations of symmetric and asymmetric windows. 6500 neural networks for window optimization were trained for a single organim group using different combinations where amino acid composition and position information were included in the input to network or not, leading to approximately 27000 neural networks being tested in all (Bendtsen et al, 2004).

It is clear that optimal signal peptide discrimination prediction requires symmetric (or nearly symmetric) windows, whereas cleavage site training needs asymmetric windows with more positions upstream of the cleavage site included in the input network. The idea of utilizing asymmetric windows in the prediction of cleavage site is due to the high variability of cleavage site position and the variability of individual amino acid residues. An asymmetric window provides a broader scanning of the cleavage site and this might reduce the prediction of false positives. The optimal window size for for cleavage site prediction for the eukaryote network included 20 positions upstream and 4 positions downstream of the cleavage site. The eukaryote discrimination network performs best when using a symmetric window of 27 positions. For both Gram-positive and Gram-negative bacteria the discrimination network is based on a symmetric window of 19 positions (Bendtsen et al, 2004). Optimal window sizes of the cleavage site network were changed by the brute-force approached used when compared to those used in siganalP 2.0 (Nielsen et al, 1997).

2.6.1.2 SignalP 3.0 performance

In evaluating the performance of SignalP version 3.0 using the same performance criteria as for the previous two versions of SignalP, performance values were calculated using fivefold cross-validation where the method was tested on sequences not present in the training data. The signal peptide discrimination performance was optimized by the introduction of a new score, termed the D score quantifying the “signal peptideness” of a given sequence segment. In the earlier versions of SignalP, the scores from the the two types of netwoeks were combined for cleavage site assignment, and not for the task of discrimination. In the the new version 3, the D-score is calculated as the average of the mean S-score and the maximal Y-score, and the two types of networks are then used for both purposes. The S-score estimates the likelihood of finding the position where a signal peptide exist while the C-score is taken to be the likelihood of the position, found from the S-score estimate, being at the +1 position with respect to the site of cleavage (Emanuelsson, 2007). The D-score is defined as the average of the maximal Y-score and

the mean S-score. It has been found that the D-score gives a better discrimination when compared to S-score or Y-score alone (Emanuelsson, 2007).
Y-score is defined as:

$$Y_i = \sqrt{(C_i \Delta_d S_i)}$$

$$\Delta_d S_i = \frac{1}{d}\left(\sum_{j=1}^{d} S_{i-j} - \sum_{j=0}^{d-1} S_{i+j} \right)$$

where $\Delta_d S_i$ is the difference between the average S-score of *d* positions after position *i* (Bendtsen, 2004).

New features regarding information about the position of the sliding window as well as information on the amino acid composition of the entire sequence to be predicted were also introduce in SignalP 3.0 to enhance its performance. It was found that information about position units were important for both the cleavage site and discrimination networks while the information on amino acid composition was found to be useful only in improving the discrimination network. The observation that the composition of secreted and non-secreted proteins differs has led to the idea of including compositional information in the improved versions of the network (Chou, 2001).

Signal peptides average lengths range from 22 (eukaryotes) and 24 (Gram-negatives) to 32 amino acid residues in Gram-positives and the new network encoding the position of the sliding window uses these averages to penalize prediction of extremely long or short signal peptides. Twin arginine signal peptides often receive a D-score below the threshold, as they tend to be quite long, averaging at about 37 amino acid residues (Palmer and Berks, 2003). The above would suggest therefore that ordinary signal peptides with extreme lengths are in a few cases not predicted by neural networks.

2.6.1.3 SignalP 3.0 comparison to other prediction methods

The data used to train a method is, in general, “easier” than genuine test sequences that are novel to a particular method (Bendtsen et al, 2004). In the SignalP version 3.0, most of the training set sequences have been retained from those used earlier by Menne et al, (2000). A common problem when comparing this tool with other methods is that they do not necessarily predict the same organism classes. The PSORT method, for example, predicts only on Gram-negative data, and not on the two other SignalP organism classes (Gardy et al, 2003).

2.6.1.3.1 PSORT

The PSORT-II method makes predictions on eukaryotic sequences, the subcellualar localization classes endoplasmic reticulum, extra cellular and Golgi were merged into one category of secretory, whereas, the rest, cytoplasmic, mitochondrial, nuclear, peroxisomal and vacuolar, were merged into a single “non-secretory” category. The reported performance was 57% accurate for all categories. As PSORT-II does not assign cleavage sites, only the discrimination performance has been compared. Most of the errors arising from the above comparison with effects on the discrimination performance owe largely to the errors in the Menne set (originating from Swiss-Prot), redundancy, as well as the first 60 residues carrying transmembrane helices in more than 10% of the novel negative test sequences (Krogh et al, 2001).

The original version of PSORT was used for predicting signal peptides in Gram-positive bacteria (Nakai and kanehisa, 1991). Outputs from “cleaved signal peptide” and “uncleaved signal peptide” have been merged into one category referred to as “secretory”. Sequences with a negative N-terminal signal peptide prediction were regarded as cytoplasmic. Again, the performance of SignalP3 was higher than PSORT. As the amount of data used to train this version of PSORT was quite small, the performance is surprisingly good (Bendtsen et al, 2004).

2.6.1.3.2 Sigfind

This is a neural network based method. When 50 randomly chosen negative and 50 randomly chosen positive test sequences from the novel Menne set were submitted to Sigfind, it returned two false positive within the negative test set and one false negative within the positive test set. When running the same sequences on SignalP3., no false positive was obtained, but the same false negative as Sigfind. Currently, Sigfind and SignalP now correctly classifies the protein as being secreted after modifying the N-terminal residues by extension of the corresponding amino acid residues (Bendtsen et al, 2004).

Sigfind correctly classifies new signal peptide-containing sequences as secretory, but classifies four of the 86 cytoplasmic and five of the nuclear sequences as secretory (false positives). SignalP 3.0 classifies all new eukaryotic secretory and cytoplasmic proteins correctly but two false positives predictions for the nuclear sequences. For discrimination of secretory and non-secretory proteins newly entered into Swiss-Prot, the Sigfind method obtains a correlation coefficient of 0.91, whereas SignalP again obtains a better correlation of 0.98. It appears that the Sigfind method quite strongly overpredicts signal peptide-containing sequences, and this means that on a normal data set (either the one used to train SignalP or a full proteome), where the non-secretory proteins greatly out number the secretory proteins, the actual performance in terms of specificity will be much lower than on this more balanced set (Bendtsen et al, 2004).

2.6.1.3.3 Comparing SignalP 3.0 to Phobius

This is a relatively new method designed to improve prediction of transmembrane helix topology by predicting topology and presence of a signal peptide and consequently integrating these predictions to yield the output prediction (Käll et al, 2004). More often than not, the first transmembrane helix can be mistaken for a signal peptide and vice versa.

Bendtsen *et al.*, (2004), made an evaluation of Phobius using the same novel sequences from release 42 of Swiss-Prot as used in the Sigfind test. Sigfind has been described above. Out of 205 negative test examples, Phobius generated two false positive predictions, whereas SignalP generated two false positivespredictions. Both methods were able to correctly classify all signal peptide-containing sequences. Correlation results showed 0.96 for Phobius and 0.98 for SignalP. Phobius could predict the position of the cleavage site correctly in 75% of the sequences, while SignalP version 3.0 is able to predict cleavage site position correctly for 87% of the sequences. For Gram-positive and Gram-negative Swiss-prot 42 sequences, Phobius had 64% accuracy of prediction compared to SignalP's 82%. However, Phobius is indeed favoured over SignalP in predicting transmembrane helices close to the N terminus, which are easier to confuse with signal peptides

2.6.1.4 SignalP misuse

In some cases, SignalP users have often interpreted a positive prediction as meaning that the protein is extracellular. This is not always the case since many proteins with signal peptides are retained at their sites of synthesis. Proteins having the endoplasmic reticulum (ER) retention signal may be retained in the ER which is found at the C terminus of the the mature protein. SignalP does not take signals such as the above into account. These wrong interpretations are even assumed to be correct in the absence of experimental data.

On the other hand, negative prediction by SignalP does not necessarily mean that the protein is indeed a non-secreted protein, as some proteins enter the extracellular space by non-classical and leaderless pathways.

2.6.2 Phobius

An inherent problem in transmembrabe (TM) protein topology prediction and signal prediction is the highly similar hydrophobic regions of a transmembrane helix and that of a signal peptide, leading to cross-reaction between the two types of predictions. TM

prediction is a classical problem in bioinformatics. Since the structure of TM proteins is difficult to determine experimentally, it is has been rewarding to predict TM topologies computationally. An alpha-helical TM segment normally consists of a 15-30 amino acid residues long region with an overrepresentation of hydrophobic residues which might seem fairly easy to recognize. However, it is complicated by the fact that many TM helices in multispanning TM proteins are partially or completely shielded by other TM helices. Since they are not completely exposed to the lipid bilayer they constitute amphipatic helices. The task to make TM topology predictions, that is, to localize all TM segments as well as determine the location (inside the cytoplasm or outside) of the loops turn out to be far from trivial (Käll *et al*., 2004).

2.6.2.1 Phobius architecture

The model architecture of Phobius can be said to be derived from joint models made within the TMHMM and SignalP-HMM where there is a transition from the last state of the signal peptide model in SignalP-HMM to the outer loop state in the TMHMM model (Käll *et al*., 2004). One of the strongest indicators of a signal peptide is a hydrophobic region which is alpha–helical in nature. It has been shown that the very popular method, SignalP V2.0.b2, is more sensitive than the other methods, and predicts cleavages more accurately, but includes many false positives (Menne *et al*., 2000).

SignalP-HMM uses a TM protein with one TM segment near the N-terminal of the protein, to help discriminate against false positives (Nielsen and Krogh, 1998), while LipoP (Juncker *et al*., 2004) is used to model N-terminal TM helices, signal peptides and lipoproteins signal peptides in gram-negative bacteria to improve discrimination between these categories. However, a joint TM topology and signal peptide predicting tool does not yet exist.

Phobius has been designed based on HMM and strives to achieve accuracy in predicting both the TM topology of a protein and the presence or absence of a signal peptide in the same protein. The ability of Phobius to discriminate TM topologies from signal peptides

gives it its main strength and its attendant accuracy of prediction as a mixed TM and signal peptides prediction tool than TM topology-only or signal peptide-only prediction tools. Several modifications have been made to the combined model of Phobius, that is, the TMHMM and SignalP-HMM. The final Phobius model, which is the architecture with the best performance, is shown in Figure 2.4 below.

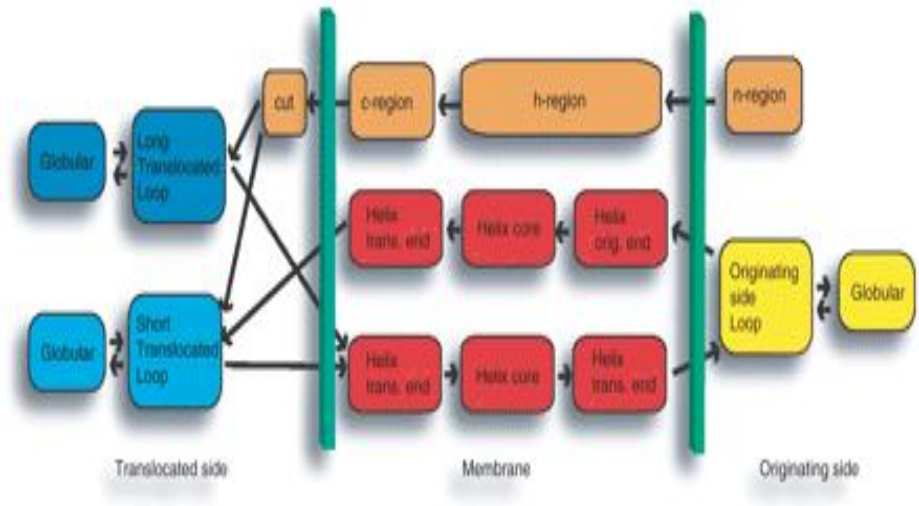

Figure 2.4 Overview of the Phobius model (Käll *et al.*, 2007)

2.6.2.2 Phobius performance

The performance of phobius has been measured by first cross-validating it and comparing it with other methods. Phobius performance has been compared to the performance of TMHMM version 2.0, HMMTOP 2.1 as well as SignalP V2.0.b2 as reported in Tables 2.3 and 2.4.

The comparisons shows that Phobius is successful in making fewer miscalculations of TM helices as signal peptides and fewer miscalculations of signal peptides as TM helices with respect to the compared methods. On the other hand, Phobius is less sensitive when predicting signal peptides and less accurate in predicting cleavage sites than SignalP, but this is well compensated for by the reduction of false positive predictions (Käll *et al.*, 2007).

Table 2.3 Accuracy of transmembrane topology predictions by Phobius, TMHMM 2.0, HMMTOP 2.1, a combination of SignalP-HMM and TMHMM, and a combination of SignalP-HMM and HMMTOP, measured for different data sets.

Proteins contain	**Both-TM and-SP**		**TM-only**		**SP-only**	**Neither-TM-nor-SP**
Test set sequences	New	All	New	All	All	All
Phobius	94.1%	91.1%	53.9%	63.6%	96.1%	98.2%
TMHMM 2.0	70.6%	71.1%	44.5%	65.2%	73.9%	98.7%
TMHMM 2.1	52.9%	51.1%	50.8%	66.8%	37.2%	85.0%
TMHMM-SignalP	88.2%	86.7%	39.5%	58.7%	98.0%	99.2%
HMMTOP-SignalP	88.2%	82.2%	45.0%	59.1%	89.6%	86.0%

TM = transmembrane, SP signal peptide (Käll *et al*., 2004).

Table 2.4 Errors in signal peptide predictions made by Phobius and SignalP V2.0.2b.

Proteins contain **Error type**	**Both-TMand-SP** **False negatives**	**TM-only** **False positives**	**SP-only** **False negatives**		**Neither-TM-nor-SP** **False positives**
Test set sequence	All	All	New	All	All
Phobius	4.4%	7.7%	2.4%	3.5%	2.3%
SignalP-NN	2.2%	42.9%	2.2%	2.3%	4.8%
SignalP-HMM	0.0%	19.0%	0.6%	1.4%	4.0%

TM = transmembrane, SP = signal peptide (Käll *et al*., 2004).

Phobius has been found to be more accurate when compared to other TM prediction methods except for the Neither-TM-nor-Signal peptide data set where it is marginally less accurate (0.5%) than TMHMM, and on the complete TM-only dataset (1.6%). The comparison also shows that HMMTOP is a better prediction method than TMHMM based on the TM-only dataset, but has obvious faults when running on data containing signal peptides or trying to sort out soluble proteins.

TM prediction methods with no regard for signal peptides can be improved by first removing any detected signal peptide by a separate signal peptide predictor before running the TM predictor (Käll *et al*., 2004). The behaviour of the type of predictor above has been reported in table 2.4 where SP detected by signalP-HMM were from the test set while TMHMM and HMMTOP were reran. Even though there was a clear increase in performance on signal peptide datasets, there was however a drop in TM-only set performance.
The training set of TMHMM is more easily predicted than the other TM sequences, an observation that is in tune with a previously drawn conclusion that the TMHMM training dataset is much easier to predict than genomic data sets (Käll *et al*., 2007). The work by Käll *et al*., (2004) indicate that SignalP-HMM is both more sensitive and selective in detecting signal peptides than SignalP_NN, but that SignalP-NN has higher accuracy in predicting cleavage sites than SignalP-HMM. The explanation for this low accuracy of Phobius has been that centered on cleavage site annotation data from Swiss-Prot when gathering the training data. It has been deduced that incorrect cleavage sites, even if just a few, have biased the model towards false sites.

2.6.3 PSORT tools

PSORT was first developed in 1990 and requests a full amino acid sequence and its sequence origin – Gram-positive bacteria, Gram-negative bacteria, yeasts, animals and

plants. At the time when PSORT was first developed, its applicability to the sequences of archea was not known (Nakai and Horton, 1999). PSORT works by selecting sequence category, e.g plant, and assigns a set of candidate localization sites. The values of feature variables that reflect various characteristics are then calculated. It then tries to interprete the set of values obtained and provides an estimated likelihood of the protein being sorted to each candidate site.

PSORT I was the first comprehensive localization prediction method to be developed for bacteria, and is capable of assigning a bacterial protein to one of several cellular compartments. The program uses a multi-component approach to prediction. Features influencing localization – including amino acid composition, sequence motifs, signal peptides and transmembrane alpha-helices – were identified in query, and the resulting information was integrated using a series of 'if-then' rules to generate a final prediction (Gardy and Brinkman, 2006) and this step was later modified through the implementation of the k- nearest neighbour classification technique (Nakai and Horton, 1999).
PSORT I had been one of the most frequently used and most popular sub cellular localization prediction tools before the development of new prediction methods around 2001. Analysis of the microbial genomes has frequently used PSORT I and it is still being used even with the availability of newer prediction methods. It, however, has its own limitations. It has recorded a low precision in its prediction power and has a high false positives rate up to 40 % (Gardy, et al, 2003) where proteins with dual localization sites, such as periphery associated membrane proteins with large cytoplasmic or periplasmic domains, are not easily identified. More so, the program fails to assign Gram-negative proteins to the extracellular space.

PSORTB is a relatively precise, multi-component approach to protein subcellular localization. It was released in 2003 and updated in 2005, and is one of the recent methods to implement the multi-component approach to prediction that was pioneered with PSORT I (Gardy and Brinkman, 2006).
PSORT-B utilizes six analytical modules in generating an overall prediction of localization site (Gardy et al, 2003). Modules included in the initial version of PSORT-B

include SCL_BLAST, PROSITE motif-based analytics, HMMTOP, a variation of SubLoc, a novel outer membrane protein analysis, and a type II secretion signal peptide prediction method.

There has been a linkage of the tendency for subcellular localization conservation to their being evolutionary in nature (Nair and Rost, 2008). PSORT-B uses SCL-BLAST for homology-based prediction while the HMMTOP is a transmembrane helix prediction tool. The identification of outer membrane proteins is of particular interest, both due to the difficulty in predicting their characteristic β-barrel structure and their high potential use as drug targets. A data mining approach was used to identify frequent sequences occurring in β-barrel proteins, both integral outer membrane proteins and auto transporter proteins, which posess a β-barrel transport domain (Gardy et al 2003). A signal peptide identification tool, and a series of frequent subsequence-based SVMs have also been included and utilized as part of the modules in PSORT-B

The modules return as output either a predicted localization site or, if the feature is not detected, a result of 'unknown'. The output is then integrated by a Bayesian network into a final prediction in the form of a score distribution that shows the likelihood of the query protein being resident at each of the nine possible localization sites (cytoplasm, cytoplasmic membrane, periplasm, outer membrane and extracellular space) and four gram-positive localizations (cytoplasmic, cytoplasmic membrane, cell wall and extracellular space (Gardy and Brinkman, 2006)).

WoLF PSORT (Horton *et al*., 2007) is basically an extension of PSORT II. The features of PSORT that are used for protein localization are also included in the design of WoLF PSORT as well as some features from iPSORT (Bannai *et al*., 2002). A weighted k-nearest neighbor method is used to classify vectors derived from the conversion of amino acids by the localization features from the PSORT tools above (Horton *et al*., 2007).

A sensitivity and specificity of approximately 70% has been recorded for WoLF PSORT for sites, such as nucleus, mitochondria, cytosol, plasma membrane, extracellular and in

chloroplasts. However, when applied to mouse proteins, WoLF PSORT was only about 50% accurate. This reduced performance in mouse when compared to the predictions in fungal and plant proteins has been linked to over-representation of well studied proteins in the training data as well as the quality and size of the test data as only 87 cytosolic proteins were in the test set(Horton *et al.*, 2007). When compared to other PSORT tools, WoLFPSORT performs very well in classifying protein subcellular localization.

2.6.3.1 Performance of PSORT tools

Protein data obtained from the second release of the bacterial protein localization database ePSORTdb was used to train and test PSORT-B (Ray et al, 2005). These proteins have been identified manually from available literature and have proven to be of high-quality as training data source. Other methods have indeed utilized this same high-quality training data sourced from literature.

When PSORT-B(v2.0) was evaluated using the above data set, it was reported that the method achieved 95.8% precision and 82.6% recall for Gram-negative proteins, and 95.9% precision and 81.3% recall for Gram-positive proteins. The application of the method to 236 bacterial genomes generated predictions for 57.2% of gram-negative proteins and 75.3% of gram-positive proteins (Gardy and Brinkman, 2006).

The advantages of PSORTB over several other methods have been highlighted. Owing to both its robust multi-component analysis and the fact that it yields an output of 'unknown if a confident prediction is not possible, it has been able to predict with a high precision'. Currently, it is the only program with the ability to automatically flag those proteins that have dual localization sites, and also the only program to give pre-computed predictions for over 200 bacterial genomes, available through the cPSORTdb (Rey *et al.*, 2005).

The accuracy of PSORT-B is a significant improvement over the PSORT I program, which according to analyses by Gardy *et al.*, (2003), had an overall precision and recall of 59.6 and 60.9% respectively, when evaluated using their data set of 1443 proteins. A

large increase in precision can be observed for each localization site; however recall is reduced in certain cases, reflecting their goal of returning an accurate prediction rather than a non-confident prediction. A 16.4% decrease in recall is compensated for by a 41.3% increase in precision for membrane proteins.

2.6.3.2 Comparison of PSORT-B v2.0 with other methods

Proteome Analyst (Lu *et al*., 2004) is capable of generating predictions for five Gram-negative localization sites and three Gram-positive sites – it does not differentiate between cell wall and extracellular proteins.

In terms of precision, PSORTB v2.0 outperforms both Proteome Analyst and CELLO by 7.6 and 26.1% respectively. The significant difference between PSORTB and CELLO is due to the fact that unlike the other two programs, CELLO forces predictions for each query protein. While this does not lead to a prediction generated for every protein in a proteome, the cost in terms of reliability of these predictions is significant. This decreased precision may not be apparent when evaluations are reported using the accuracy measure, in which high recall is able to compensate for lower precision, and illustrates that reporting confusion matrices leads to, epigrammatically enough, the least confusion when comparing the performance of multiple programs (Gardy *et al*, 2005).

When PSORTB v2.0 was used to analyze the same organisms earlier analyzed by Proteome Analyst, it attained coverage of 68.1% for *Pseudomonas aeruginosa* and 76.5% for *Bacillus subtilis*. Proteome analyst had displayed coverage of 75.6% for the Gram-negative bacterium *P. aeruginosa* and 67.2% for the Gram-positive bacterium *B. subtilis*. Analysis based on these two proteomes suggests that while Proteome Analyst attains higher coverage on a Gram-negative organism, PSORTB v2.0 generates more predictions for a Gram-positive proteome (Gardy *et al*, 2005).

2.6.4 LocateP

Among the Sec-dependent exported proteins, current predictors have difficulties in distinguishing proteins that are cleaved from the cell membrane by the corresponding

cleavage enzyme – the type I signal peptidase. LocateP is geared towards identifying the detailed subcellular localization of bacterial proteins by combining a set of existing prediction methods while a special effort has been made to raise prediction accuracy of the N-anchored proteins.

2.6.4.1 LocateP pipeline

The LocateP pipeline was designed such that it mimics the protein secretion process in Gram-positive bacteria. The pipeline structure can be categorized as follows: (1) secretion pathway prediction, (2) transmembrane-segment detection, (3) signal peptide identification, and (4) cleavage and retention signal recognition. The LocateP pipeline employs existing subcellular localization prediction tools as well as new and more accurate methods developed by Zhou et al, (2008), for the prediction of lipoproteins, TAT-secreted, N-terminally anchored, C-terminally anchored and secreted proteins. LocateP uses at least two prediction methods for each subcellular location, in order to increase accuracy the selection of which was derived from literature. The LocateP pipeline is captured in Figure 2.5.

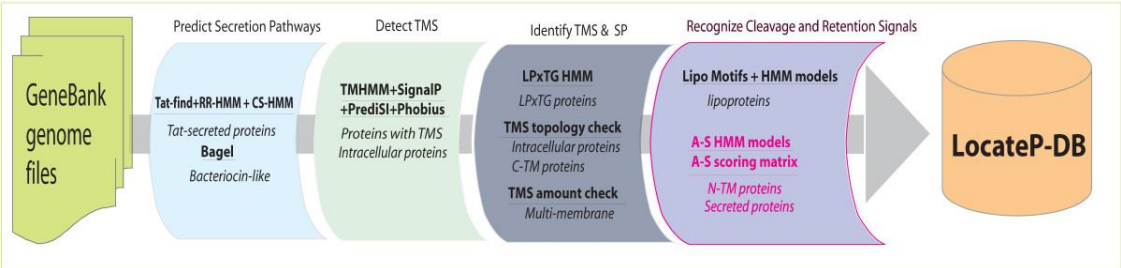

Figure 2.5 The LocateP pipeline (Zhou *et al*., 2008).

A good number of membrane proteins which possess a single N-terminal TM anchor are easily identified due to absence of a predicted cleavage site for signal peptidases even though the predictions of subcellular location of proteins containing a putative signal peptidase type I cleavage site can be particularly difficult for recent subcellular

localization prediction tools. While Sec-exported proteins are cleaved by signal peptidase, a considerable number of proteins remain membrane-anchored via the N-terminus (Djalsma and Dijl, 2005).

To identify the features that determine cleavage, multiple sequence aligments of the signal peptides from experimentally validated N-anchored and secreted proteins containing a putative signal peptidase cleavage site in *Bacillus subtilis* were analyzed. A series of HMMs were constructed based on the sequence alignments of the N-anchored and secreted proteins. Different numbers of residues on either side of the putative cleavage site were included in the models in order to investigate the roles of the H-region and C-region in cleavage-site recognition (Zhou *et al*., 2008).

2.6.4.2 LocateP validation and performance

The experimental data of Tjalsma *et al*., (2005), which is proteomic data derived from novel insights into the mechanisms of protein export from *B.subtilis*, has been used to test LocateP leading to a construction of an HMM pair capable of distinguishing the N-anchored and secreted protein containing a putative signal peptidase I cleavage site. It was found that LocateP was able to distinguish these proteins with an accuracy greater than 90%. Subsequently, LocateP was tested on ten other data sets extracted from literature describing other sub cellular localization prediction tools, and it performed extremely well on these sets with prediction accuracy higher than 90% (Zhou *et al*., 2008).

TransportDB (Ren *et al*., 2004) provided data used for a second check on LocateP performance where 1336 transport-related proteins from *Bacillus subtilis* 168, *Bacillus cereus* ATCC14579 and *Lactobacillus plantarum* WCFS1 were verified. LocateP was able to identify 113 of 124 proteins, for the substrate binding of ABC transport systems, in a correct subcellular localization for substrate binding; 96 lipoproteins, 6 secreted and 11 N-anchored proteins (Miethke *et al*., 2006, Ollinger *et al*., 2006).

A third and less quantitative check included a comparison of the LocateP predictions for all N-anchored and secreted proteins of the *Bacillus subtilis* genome with their NCBI functional annotations. Nearly all of the predictions appeared to make biological sense according to literature – most of the predicted N-anchored proteins were annotated to be involved in processes that are related to the cell-envelope, such as cell divison, cell-envelope biogenesis, etc (Zhou *et al*., 2008). Predicted secreted proteins tended to be defined to be secreted enzymes, mainly alkaline phosphatases, metalloproteases, neutral proteases, and subtilisin-family proteases.

2.6.4.3 Comparison of LocateP with other methods

The performance of LocateP was compared with TatP and Tatfind v1.2 on the proteins containing a RR/RK pattern in their N-terminus. LocateP clearly performed better than the other two specific tools when tested with intracellular and membrane protein sets via the TAT-pathway in 22 species that apparently lacked the revelant genes (Dilks *et al*., 2003), whereas LocateP did not find any Tat-pathway substrates in those species. Thus LocateP showed the best overall performance among the Tat-pathway prediction tools for gram positive bacteria (Zhou *et al*., 2008).

2.6.5 SubLoc

SubLoc was the first publicly available SVM-based localization tool to be released, generating predictions through the analysis of the overall amino-acid composition of a protein (Gardy and Brinkman, 2006). SVM method has been described earlier. The SVM approach in SubLoc is very well founded theoretically because it is based on extremely well developed machine learning theory, Statistical Learning Theory (Vapnik, 1998) as well as having superior practical applications.

2.6.5.1 SubLoc design and Implementation

Protein subcellular localization prediction is a multi-class classification problem. For SubLoc, the class number is equal to 3 for prokaryotic sequences and 4 for eukaryotic

sequences. A simple strategy to handle the multi-class classification is to reduce the multi-classification to a series of binary classifications. For a k-class classification, k SVMs are constructed. The *i*th SVM will be trained with all of the samples in the ith class with positive labels and all other samples with negative labels. The algorithm spends less time on the classification of unknown samples because we only need to calculate the inner products between the unknown samples and a small subset made up of the support vectors making SVM an efficient classifier (Hua and Su, 2001).

2.6.5.2 SubLoc performance

SubLoc was trained and tested on the Reinhardt and Hubbard dataset which was from the SwissProt database. An overall recall of 91.4% and Matthews Correlation Coefficient (MCCs) ranging from 0.68 to 0.86 has been reported for each localization (Gardy and Brinkman, 2006).

Prediction accuracies by jackknife tests for prokaryotic sequences suggest that the current method reached an accuracy of 89.3% with the simplest kernel function. This indicates that the prokaryotic samples can be well separated by a proper linear hyperplane in the input space. The accuracy could be improved by using the more complex non-linear kernel, in which the accuracy was improved to 91.4% (Hua and Su, 2001).

For eukaryotic sequences, the training procedure did not converge when a linear kernel was used which suggested that no hyperplane in the input space can clearly separate the eukaryotic samples.

2.6.5.3 Comparison of SubLoc with other methods

In comparing SVM method predictions with other methods, the Reinhardt and Hubbard dataset was also tested with the neural network method (Reinhardt and Hubbard, 1998) and the covariant discriminant algorithm (Chou and Elrod, 1999). The methods above were based only on amino acid composition. The jack knife test was used to obtain results from Markov model, covariant discrimination, and the SVM method while the neural network method results were with 6-fold cross validation.

The total accuracy of the SVM method is approximately 10% higher than that of the neural network and higher than that of the covariant discriminant algorithm by about 5% for prokaryotic sequences. For cytoplasmic sequences, accuracy reached 97.5% with the SVM method when compared with other methods. For eukaryotic sequences, the total accuracy was 13% higher than that of the neural network method where also the prediction accuracies for nuclear and cytoplasmic sequences were 15% and 22% higher than those of the neural network method although the accuracy for mitochondrial sequences were about 40% lower, indicating a room for improvement (Hua and Su, 2001).
When compared with Markov chain model, which was based on the full sequence information including the order information, the total accuracies using the SVM method were 2.3% higher for prokaryotic sequences and 6.4% higher for eukaryotic sequences.

Although the release of SubLoc marked an important milestone in the use of SVMs for localization prediction, the method itself has two significant caveats. Known and suspected cytoplasmic membrane and outer membrane proteins must be removed before the analysis since the program sorts proteins to three locations (cytoplasm, periplasm and extracellular space). Furthermore, the method does not distinguish between Gram-positive and Gram-negative queries, therefore proteins from Gram-positive organism can be mistakenly classified as periplasmic (Gardy and Brinkman, 2006) .

Table 2.5 summarizes the performance of most of the tools used in predicting protein secretory signals. They have been categorized into those that identify exported proteins, cytoplasmic membranes proteins, and outer membrane proteins. Phobius appears to me to be the best for predicting exported proteins but has too many FP predictions when predicting cytoplasmic membrane proteins. Proteome Analysist appears to be the best for predicting both cytoplasmic and outer membrane proteins even though it returns FP results when predicting exported proteins.

Table 2.5 Performance of feature-based detection methods versus general localization method.

Program	**Performance Statistics**					
	TP	FP	FN	TN	Precision	Recall
Identification of exported proteins proteins						
Phobius	125	4	29	141	96.9%	81.2%
PSORTb	110	4	44	141	96.5%	71.4%
CELLO	128	10	26	135	92.8%	83.1%
P-CLASSIFIER	125	10	29	135	92.6%	81.2%
Proteome Analyst	119	10	35	135	92.2%	77.3%
LOCtree	99	8	55	137	92.5%	64.3%
PSLpred	129	11	25	134	92.1%	83.8%
SignalP,LipoP,TatP,TMHMM	126	12	28	133	91.3%	81.8%
SubLoc	86	10	68	135	89.6%	55.8%
PSORT I	129	34	25	111	79.1%	83.8%
Identification of cytoplasmic membrane proteins						
Proteome Analyst	59	2	10	228	96.7%	85.5%
PSORTb	55	2	14	228	96.5%	79.7%
CELLO	43	2	26	228	95.6%	62.3%
P-CLASSIFIER	41	2	28	228	95.3%	59.4%
Phobius	54	3	15	227	94.7%	78.3%
PSLpres	48	8	21	222	85.7%	69.6%
TMHMM	53	12	16	218	81.5%	76.8%
ConPredII	56	20	13	210	73.7%	81.2%
PSORT I	54	39	15	191	58.1%	78.3%
Identification of outer-membrane proteins						
Proteome Analyst	31	0	7	261	100.0%	81.6%
PSORTb	30	0	8	261	100.0%	78.9%
PSORT I	33	4	5	257	89.2%	86.8%
PSLpred	21	3	17	258	87.5%	55.3%
BOMP	20	4	18	257	83.3%	52.6%
Prof-TMB	19	7	19	254	73.1%	50.0%
TMB-Hunt	18	7	20	254	72.0%	47.4%
CELLO	21	10	17	251	67.7%	55.3%
P-CLASSIFIER	20	13	18	248	60.6%	52.6%

FN =False negative; FP = false Positive; TN = true negative; TP = true positive (adapted from Gardy et al., 2006)

3. Biological motivation for the study

The secretion of proteins across biological membranes is in most cases mediated by translocation machinery recognizing a specific sequence tag or motif in the protein to be secreted. In bacteria, the classical tripartite structured Sec signal peptide governs most of the targeting to the secretion pathway. There are also various pathways which have been discovered to work in a Sec-independent fashion, most predominant of which is the twin-arginine translocation (Tat) secretion pathway where a twin-arginine consensus motif is located within the signal peptide itself. The N-terminal plays a central role in these secretory systems as the tag signaling secretion (Bendtsen et al, 2005). Bacterial proteins have been found to be secreted without an apparent signal peptide. This phenomenon termed non-classical secretion was identified many years ago when interleukin 1β and thioredoxin were discovered to be secreted without any identifiable signal peptide (Rubartelli *et al*., 1992).

Some proteins, which have been found to display a function in the cytoplasm, have also been shown to actively participate in biological process in the extracellular environment, but this does not necessarily imply that their function in the extracellular environment is identical to that in the cytoplasmic environment. Such proteins which display two unrelated functions have been so named “moonlighting” proteins (Schaumburg *et al*., 2004; Jeffery, 2003). A basic strategy to identifying non-classically secreted proteins is by inactivating the Sec-dependent pathway by mutation or chemical treatment. Hirose et al, (2000), used SecA mutants to disrupt the translocation machinery, allowing them to identify several non-classically secreted proteins in *B.subtilis*.

Examples of non-classical secretion using a set of five named bacteria proteins are presented here. Studies of non-classical secretion in bacteria have been carried out on the human pathogen *Mycobacterium tuberculosis*, and it was found that this organism secretes glutamine synthetase (GlnA). The GlnA in the non-pathogenetic *Mycobacterium smegmatis* is found solely in the cytoplasm. When a recombinant GlnA from *M.*

tuberclosis was expressed in *M. smegmatis*, it was found to be secreted also, suggesting that the signal for secretion is contained within the protein sequence (Harth and Horwitz, 1997). Glutamine synthetase is involved in purine biosynthesis. Specifically, it catalyzes the reaction of glutamate and ammonium ions to yield glutamine. Glutamine synthase is found in all organisms, and in addition to its importance for ammonium ions assimilation in bacteria, it has a central role in detoxification in mammalian blood (Nelson and Cox, 2005).

Enolase is involved in catalyzing a step of glycolysis, where 2-phosphoglycerate is dehydrated reversibly. Metal ion catalysis and transition state stabilization is illustrated by the enolase reaction (Nelson and Cox, 2005). The covalent binding to the substrate causes inactivation of the enzyme, and possibly serves as a signal for the export of the protein. Proteins required in the metabolism of carbohydrate, such as, Eno, pdhB, PdhD and CitH, were identified as being extracellular by Vitikainen et al, (2004), but it was discovered that none of these proteins had a known signal peptide.

GroEL belongs to the chaperonin family of molecular chaperones, and is found in large amounts in bacteria. It is required for the proper folding of many proteins. Within the cell, the process of GroEL/ES mediated protein folding involves multiple rounds of encapsulation, and release of substrate protein. Structurally, GroEL is a dual-ringed tetradecamer, with both the cis and trans rings consisting of seven subunits each. The inside of GroEL is hydrophobic, and probably accounts for the location where protein folding takes place (Horwich *et al*., 2007).

GroEL has been found to be secreted to the extracellular space as well as on the cell surface. Both elongation factor EF-G and the protein folding chaperon, GroEL have been found in the extracellular environment from studies carried out by Antelmann *et al*., (2001) and Vitikainen *et al*., (2004). However, these two proteins were absent in the Sec deletion analysis carried out by Hirose *et al*., (2000). Studies on the GroEL homologue HspB has been shown to be actively secreted in stationary phase *Helicobacter pylori* even though GroEL is known to have a cytoplasmic function (Vanet and Labigne, 1998).

During the elongation phase of protein synthesis, each aminoacyl-tRNA (aa-tRNA) is transported to the mRNA-programmed ribosome as a ternary complex with elongation factor Tu (EF-Tu) and GTP. On binding to the ribosome, aa-tRNA enters the A/T state, where it occupies the A site on the 30S ribosomal subunit while still interacting with EF-Tu. In the presence of a codon–anticodon interaction, GTPase activity of EF-Tu is greatly stimulated and GTP hydrolysis occurs rapidly. This induces a large conformational change of EF-Tu followed by its dissociation from the ribosome and subsequent accommodation of aa-tRNA into the ribosomal A site (Villa *et al*., 2009). The elongation factor EF-Tu has been shown, from studies carried out by Schaumburg *et al*., (2004) and Lenz *et al*., (2003), to be found on the cell surface of the Gram-positive pathogen, *Listeria monocytogenes*.

Another example of a "non-classically" secreted protein is arginase. Arginase catalyzes the hydrolysis of L-arginine to yield L-ornithine and urea. In humans there are two isozymes, arginase I and arginase II, that share ~60% sequence identity. Ornithine is the biosynthetic precursor of polyamines that facilitate cellular proliferation and tumor growth. Additionally, arginase I is implicated in tumoral immune evasion, and arginase inhibitors therefore have significant chemotherapeutic potential (Constanzo *et al*., 2007). Antelman *et al*., (2001) initially found two proteins, RocA and RocF, involved in amino acid metabolism to be non-classically secreted. However, RocF was later identified by Vitikainen *et al*., (2004) to be secreted. The RocF gene is known to code for the enzyme, arginase. Arginase was found to be secreted extracellularly, but the precise extracellular function has not been reported.

This work seeks to test the hypothesis that proteins secreted via non-classical routes could have the wrong conceptual translation (wrong translation start sites (TSS)) or they might be using a novel signal peptide. In fact wrong start sites are the most common mistake a gene prediction algorithm can make. The strategy here is to examine an alternative gene prediction approach directed at optimizing the start site. To this end, we

have gone a step further by starting from the DNA level in these alternative prediction methods while trying to reduce TSS effects. It is currently unknown whether secretion by non-classical means occur at a specifically localized membrane as seen for secretion of SpeB in *Streptococcus pyrogenes*. Indeed, the mechanism or mechanisms responsible for non-classical secretion are unknown (Bendtsen *et al.*, 2005).

4. Methods

In this chapter, the methods used in this research are presented.

4.1 Data

Protein sequences were obtained from NCBI with the following gene IDs shown in table 4.1 below. Enolase, GroEL, EF-Tu,and arginase were from *Lactobacillus acidophilus* NCFM while glutamine synthase was from *Bacillus subtilis*. The annotated protein for Lactobacillus acidophilus glutamine synthase did not match its supposed corresponding annotated gene after translation. This tends to suggest some kind of annotation error and hence glutamine synthase from another organism was used.

Table 4.1 Source of protein data

Protein	**Gene ID**	**Organism**
Arginase	3251201	*Lactobacillus acidophilus NCFM*
Enolase	3251804	*L. acidophilus NCFM*
EF-Tu	3251716	*L. acidophilus NCFM*
GroEL	3252952	*L. acidophilus NCFM*
Glutamine synthase	936028	*B. subtilis 168*

4.2 Start site optimization

More often than not, data for everyday bioinformatic research come from existing databases, some of which have been wrongly annotated. In predicting secretory proteins, by way of looking out for signal peptides, attention is usually paid to the correctness of the start site in the annotated protein. Naturally, if a protein has a wrongly assigned start

site, it can lead to other false predictions, such as that of the signal peptide. Part of the goal of this work is to utilize sequences whose start sites have been optimized as far as possible.

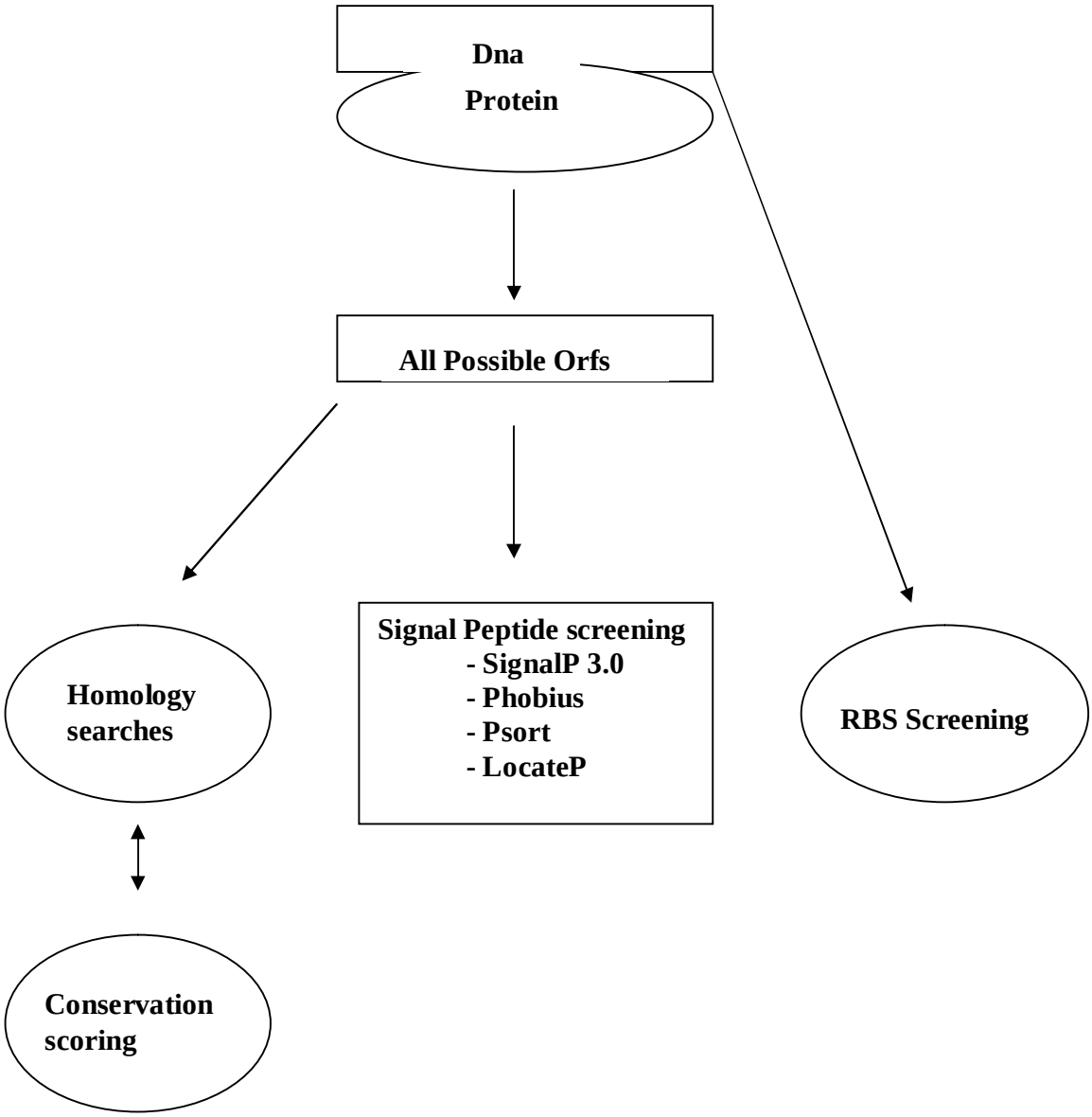

Figure 4.2 A flow chart of analysis pipeline. Longest orf was used for homology searches.

4.2.1 Extension of DNA flanks

The genes coding for the test proteins were obtained and extended by 500 base pairs upstream and downstream. This was done fairly easily by simply using the “update” button in NCBI. It should be noted here that the choice of extending the starting gene by 500 base pairs is not rigid. The idea was to have a relatively meaningful number of nucleotides before the original annotated start and end positions since we are interested in checking for alternative start sites in the now extended DNA.

4.2.2 Retrieval of open reading frames

All possible open reading frames were retrieved by implementing a perl script and confirmed manually. The script takes as its input the extended DNA sequence, the original protein sequence, and a codon table file. Here, the bacterial codon table was used. Since all possible orfs were considered, the codons ATG, TTG, CTG, ATT, ATC, ATA, and GTG were all translated to the corresponding methionine. The script first reads the DNA in all six frames, translates the DNAs into proteins, and returns only the protein sequence which matches the original sequence. The script then splits the protein sequence into all possible open reading frames with the longest orf being the first and the shortest orf being the last as splitting occurs from the N-terminal end of the protein.

4.3 Signal peptide screening

The orfs, as well as the original protein sequence, were screened for signal peptides using SignalP 3.0 (Bendtsen, 2004), Phobius (Käll *et al*., 2004), PSORT (Nakai and Horton, 1999), and LocateP (Zhou *et al*., 2008). The choice of these tools for finding signal peptides was informed by their relative performances. SignalP 3.0 combines the HMM as well as the NN approach in its prediction. This coupled with improved scoring methods

by its recently introduced D-score makes it even more reliable. Phobius, on the other hand, has been shown to perform very well when compared with other methods. Phobius also has the advantage of predicting a TM topology alongside a signal peptide. Even though PSORT is relatively old by virtue of its release, it is still very good at predicting SCLs for bacteria especially. In fact, PSORT was actually developed for use in bacteria.

4.4 Ribosomal binding site screening

The presence or absence of a ribosomal binding site was checked with Patser v3e (van Helden, 2003). The program requires a DNA sequence, a motif file, and an alphabet file. In this case the motif file used was a matrix file for the motif in question. Here, matrices for two very general RBS motifs were used as shown in table 5.5 in the next chapter. The matrices were constructed from a multiple sequence alignments of sequences where the ribosomal site has been determined or at least known to exist. The frequency of the nucleotides constituting the RBS motif is used to build the corresponding matrix. During protein synthesis, the ribosome must bind to an mRNA at some defined and recognizable site for translation to occur. Screening for this site is informed by the fact that a correct prediction of the ribosomal binding site will also give the indication of having predicted a correct start site. The score values as well as the natural logarithm of the P-value were used to discriminate against possible RBSs found elsewhere in the DNA. In fact only positions from the highest scores were reported.

4.5 Homology searches/evaluation of conservation of start sites

Using NCBI BLAST, the longest orf from all possible orfs retrieved from each of the proteins above was used to search for homologues. Multiple sequence alignment (MSA) was done for the 20 best hits for each protein using MUSCLE (Edgar, 2004) alignment tool. The MSA output was exported as an *.aln* file for conservation analysis.

Conservation scoring was calculating using the Shannon entropy method (Capra and Singh, 2007). Shannon entropy (SE) is one of the simplest and most common measures of conservation at a site. It is defined for a column *C* as:

$$\mathrm{SE}_C = -\sum_{\alpha \in AA} p_C(\alpha) \log p_C(\alpha)$$

where $p_C(\alpha)$ is the probability of the individual amino acid in *C* (Capra and Singh, 2007).

Multiple sequence alignments were also visualized using GeneDoc (Nicholas *et al*., 1997) and exported to word document.

5. Results

This chapter presents the results of the experiments carried out above. The results from the homology searches/ conservation analysis, evaluation of the presence of signal peptides in the protein sequences as well as the screening for ribosomal binding site in the extended DNA are presented here. An attempt has been made here to predict actual secretory signals by strategically combining results from all of the applicable methods used in the analysis.

5.1 Detecting signal peptides and/or cleavage site with SignalP 3.0

A total of 305 orfs derived from all 5 original protein sequences, which are representative examples of non-classically secreted proteins, were screened using SignalP 3.0. 53 of these were from arginase, while 50, 64, 58, and 80 were from enolase, GroEL, EF-Tu, and glutamine synthase respectively. Orfs were numbered serially as subsequences beginning from the N-terminal end of the original (starting) protein from 1 to … n, where n = last orf.

A close look at the output scores from the SignalP analysis above reveals a seemingly uniform low HMM scores which can be assumed to be insignificant. For the purpose of further analysis and comparison, only significant HMMs will be used along D-scores. The results obtained show that prediction of signal peptides by SignalP 3.0 is enhanced in the protein sequences that have had their start sites optimized when compared with predictions from the original protein sequences as shown in figure 5.1 below

Table 5.1 SignalP 3.0 best's scores

Protein	**OS D-score**	**OS HMM-Score**	**NS D-score**	**NS HMM-score**
Arginase	0.324	0.000	0.128	0.000
Enolase	0.411	0.032	0.118	0.000
EF-Tu	0.306	0.000	0.143	0.000
GroEL	0.621	0.678	0.176	0.000
Glutamine Synthase	0.225	0.001	0.069	0.000

OS = Optimized sequence, NS = Original sequence. The D-score and HMM-score are derived from the SignalP-NN and SignalP-HMM. While the D-score is used to discriminate between SPs and non-SPs for a given cut-off value, HMM-score is based on posterior probabilities.

Table 5.2 SignalP 3.0 best scores at corresponding orfs

Protein	**Orf**	**OS D-score**	**OS HMM-score**
Arginase	16	0.324	0.000
Enolase	10	0.411	0.032
EF-Tu	48	0.306	0.000
GroEL	59	0.621	0.678
Glutamine Synthase	36	0.225	0.001

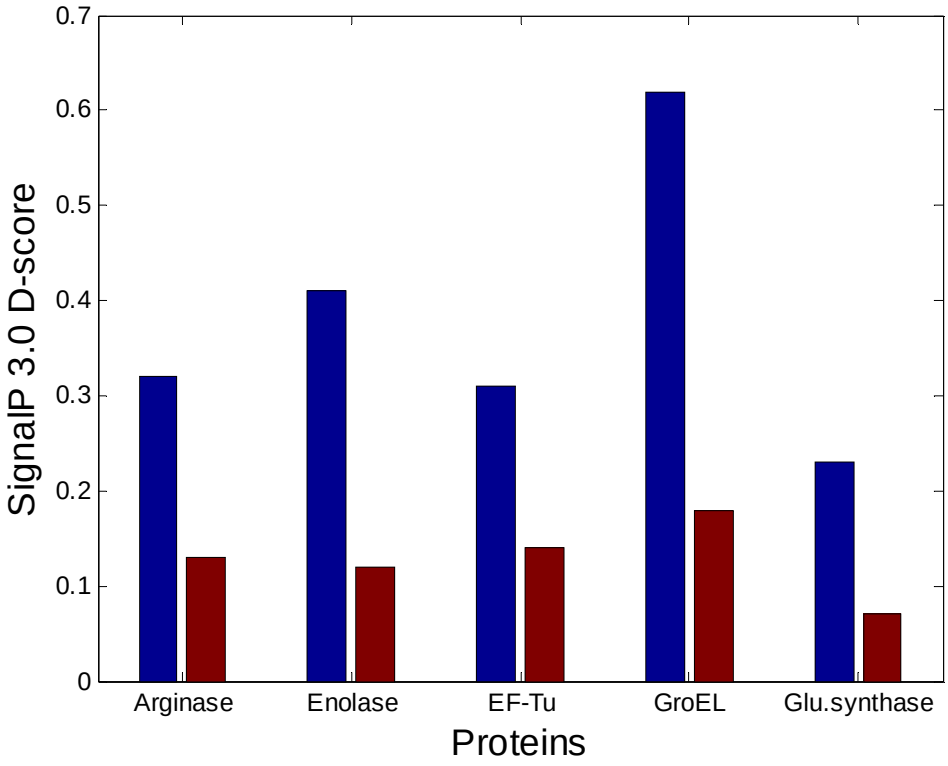

Figure 5.1 SignalP 3.0 Prediction in Optimized and Original Protein Sequences. Optimized proteins sequences correspond to blue bars while red bars represent original protein sequences.

When compared to the D-score cut-off value of 0.45, it was found that the SignalP 3.0 actually predicted a signal peptide/cleavage site in GroEL. The prediction of a best score of 0.411 compared to the cut-off value of 0.45 also offers interesting insights into the possibility of an actual signal peptide prediction in enolase. More light will be shed on this when a combined prediction is applied in the later part of this chapter. This also shows the trend already obtained above where predictions in the optimized protein sequences were seen to yield a better prospect for finding secretoty signals. Figure 5.2 is

a plot showing the predictions in optimized protein sequences and the original sequences with respect to the D-score cut-off.

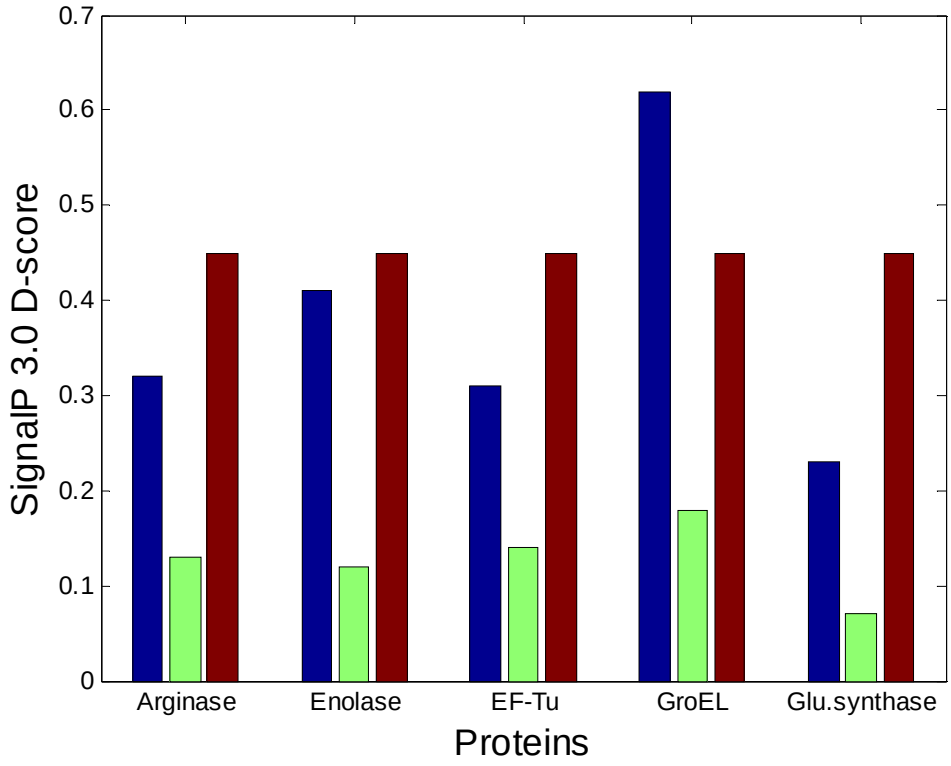

Figure 5.2 Comparison of SignalP predictions in Optimized and original protein sequence with reference to prediction cut-off score. Optimized proteins sequences correspond to blue bars while green bars represent original protein sequences. D-score cut-off is shown in red bars.

5.2 Detection of signal peptides from alternative gene predictions could be by chance

When predicting signal peptide sequences by trying out multiple start sites (alternative gene prediction) there is always a question of increasing predictions just by chance. To

see how unique the predictions made are, the best D-scores for SignalP have been listed as shown Table 5.3. Here we consider GroEL where an actual prediction was made, and enolase whose best score of 0.41 has been compared to the D-score cut-off of 0.45. The idea is to check uniqueness of these scores by looking at the gap between the best scores and other scores made by the predictor. For example, GroEL's second best score also implies that there is a signal peptide present in that orf in GroEL and even the next high scoring orf (0.44) is as close as the D-score cut-off.

Table 5.3 SignalP 3.0 top D-scores.

OS D-score	
GroEL	**Enolase**
0.62	0.41
0.59	0.40
0.44	0.38
0.39	0.36
0.31	0.32
0.30	0.31

OS = optimized sequence

Since signal peptides are unique, and it is very unlikely to find many signal peptides in one protein sequence, it is therefore important to check their 'uniqueness' especially when using the kind of methodology employed here. The properties of signal peptides are unique and probably do not occur in random sequences.

5.3 Secretory signal predictions from Phobius and PSORT

An evaluation of the presence of secretory signals of all orfs using Phobius reveals the presence of a signal peptide and a transmembrane element in enolase as well as a signal peptide in GroEL. In enolase, transmembrane was found in orf9 while the signal peptide

was found in orf 10. Phobius also detected signal peptides in orfs 59 and 60 of GroEL respectively. The prediction of secretory signals by PSORT has been summarized in Table 5.4. Orfs in which extra cellular predictions were made are enclosed in parentheses in the table.

Table 5.4 Summary of PSORT predictions.

Protein	**Total number of orfs**	**PSORT predictions**			
		Cytoplasmic	Unknown	Extracellular	Multiple sites
Arginase	53	29	15	9(orfs 45-53)	-
Enolase	50	49	-	1(orf 50)	-
EF-Tu	58	48	-	5(orfs 54-58)	4
GroEL	64	57	-	5(orfs 60-64)	2
Glutamine synthase	80	64	8	8(orfs 73-80)	-

For each optimized protein sequence, the number of orf(s) for which subcellular localization prediction has been made is shown in the table. In the case of extracellular predictions, the specific orfs involved are listed serially where they are more than one.

5.4 No significant detection of Secretory signal from LocateP

LocateP results as presented in Table 5.5 did not show any predictons for a signal peptide in all the protein sequences tested. It should be stated here that only original protein sequences were screened with LocateP for obvious reasons – it stores the results of pre-determinant proteins of whole genomes where the name of the particular protein of interest is used to search the database for results.

Table 5.5 Predictions from LocateP.

Protein	**LocateP predictions**					
	LP	SWP	PP	IP	SP	CS
Arginase	No	Cytoplamic	Intracellular	1.0	-1.0	No
Enolase	No	Cytoplasmic	Intracellular	1.0	-1.0	No
EF-Tu	No	Cytoplasmic	Intracellular	1.0	-1.0	No
GroEL	No	Cytoplasmic	Intracellular	1.0	-1.0	No
Glut. synthase	No	Cytoplasmic	Intracellular	1.0	-1.0	No

Predictions were made only on non-optimized (original) sequences. LP=localization pathway, SWP= SwissProt, PP = pathway prediction, IP = intracellular possibility, SP = signal peptide, CS = cleavage site

5.5 Ribosomal binding site (RBS) screening offers useful results

Results from RBS finding allows us to compare some of the results already obtained above by looking at the position(s) on the DNA. The positions used here correspond to those having the best scores as shown in Table 5.6. In cases where more than one RBS position is found for a protein, it may imply a possibility of the ribosome binding at any of the positions found. Here, the scores from each motif that has been used for screening is first considered. If two positions are found with the best score, it is easy to assume the ribosome binding to either of the positions.

As shown in Table 5.6 below, the p-values are extremely important for checking significance of results obtained from analysis of data. Smaller p-values tend to give the results obtained some level of significance so that one is able to decide fairly easily whether the results are to be disregarded or accepted. In the case of ribosomal binding site (RBS) screening, many different positions are returned from searches when using different motif matrices. Sometimes, one orf may in fact have more than one RBS site returned for it as can be seen in Table 5.6. Since p-values are usually very small fractions/percentages, obtaining the natural logarithm of p-values yields negative values which are easier to visualize and interpret. If a RBS has been found in more than one

position in a particular orf, the position with the smaller ln(p-value) is usually taken to be more reliable. The ln(p-value) gives a certain level of confidence in choosing the results that one can rely on. However, the fact that ln(p-value) is useful in assessing the reliability of the results obtained does not mean that significant results are necessarily the correct or desired

Table 5.6 Summary of results from RBS screening

Protein	**Position(s)**	**Score**	**ln(p-value)**
Screening with AGGAGG matrix			
Arginase	500	4.98	-8.32
Enolase	1263, 1494	4.69	-7.62
EF-Tu	687, 1065, 2092	4.69	-7.62
GroEL	142, 500	4.98	-8.32
Glutamine synthase	652	4.69	-7.62
Screening with AAGGAG matrix			
Arginase	501	4.85	-6.71
Enolase	501	6.44	-8.32
EF-Tu	501	6.44	-8.32
GroEL	143	5.02	-7.22
Glutamine synthase	482	4.85	-6.71

5.6 Homology and conservation analysis offers useful insights

The results from conservation analysis are summarized below in Table 5.7. A smaller entropy score generally indicates a better conservation as reflected in the protein sequences below. Since the scores in the table were derived from the average of 10 scores before and after the start site of the original (non-optimized protein), it gives us an idea of how conserved the start is. It is easier to make prediction when we can get some idea of "rigid" the area around the start site can be or how "flexible" otherwise. GeneDoc

visualizations for the first part multiple sequence alignment of the protein sequences from the N-terminal are shown in figures 5.3 (A-E) for emphasis.

Table 5.7 Conservation score results from multiple sequence alignments.

Protein	**Conservation Score**
Arginase	0.3996
Enolase	0.7406
EF-Tu	0.5152
GroEL	0.9164
Glutamine synthase	0.4046

Conservation scores are average of +/- 10 residues on both sites of start site.

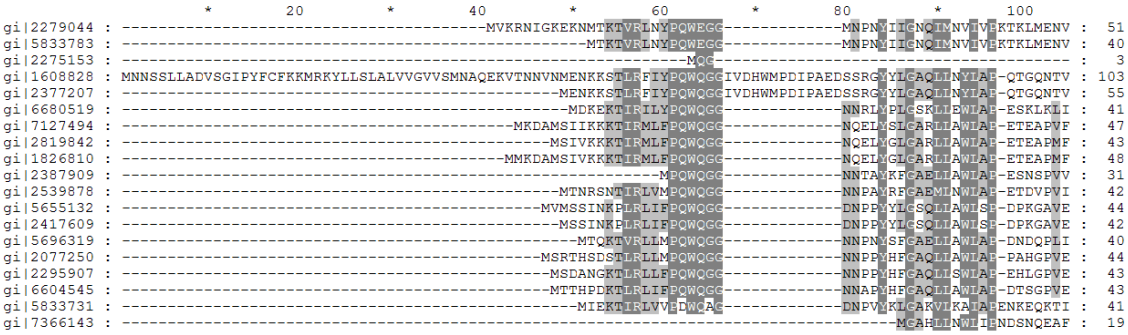

Figure 5.3(A) Arginase

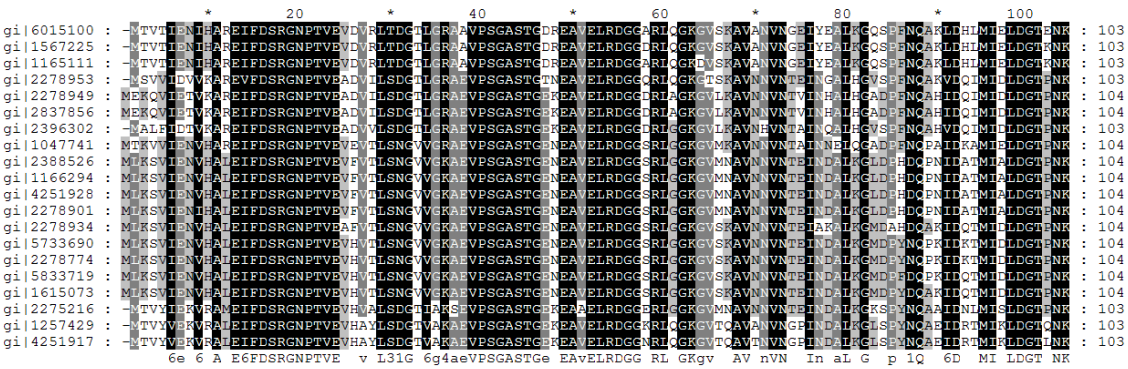

Figure 5.3(B) Enolase

Figure 5.3(C) GroEL

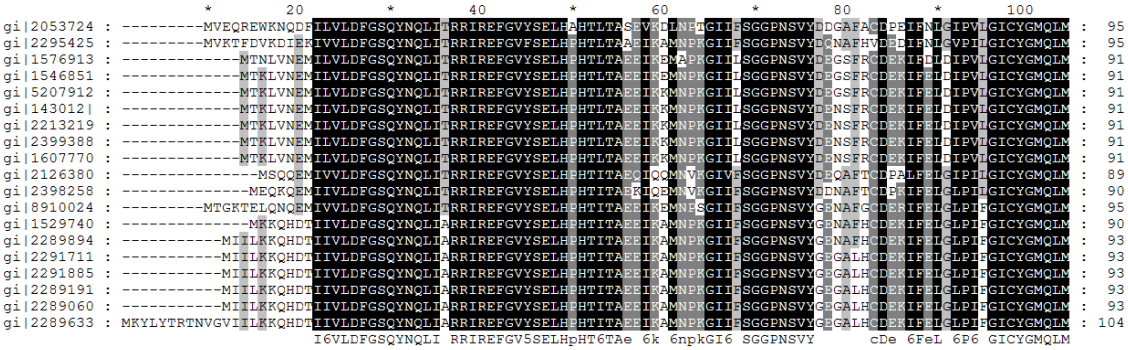

Figure 5.3(D) Glutamine synthase

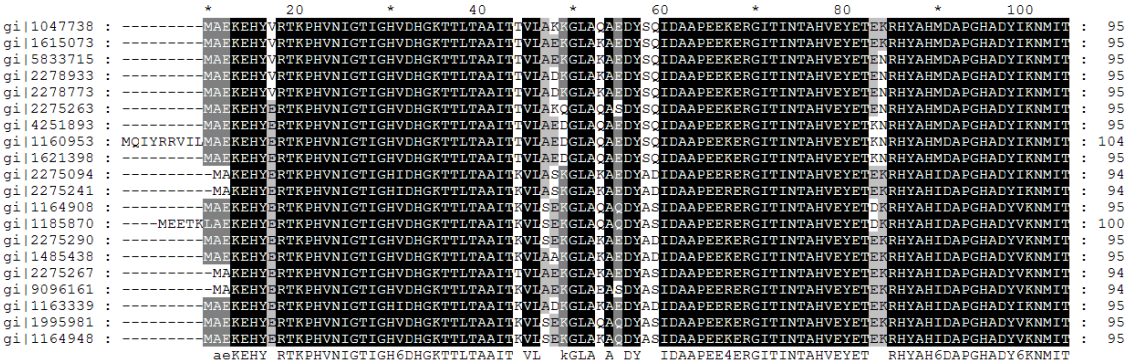

Figure 5.3(E) EF-Tu

Figures 5.3 (A-E) Multiple sequence alignments from homology searches showing part of the sequences for arginase, enolase, GroEL, glutamine synthase, and EF-Tu respectively. The alignments around the start sites are important for assessing how well the start site has been predicted.

6. Discussion

In this chapter, the methods used in this work are discussed while highlighting the importance of data in the development of some of the tools used here as well as other well established tools. The results of this work are then discussed, and an outline of the prospects and future developments of this project is suggested.

6.1 Methods of data analysis

In the course of this work, several methods have been used. These methods have been developed for different tasks and by utilizing a varied array of data for training or testing depending on the type of program. Within the period of initial review of literatures for this work, it became very evident that quality of data is a key element when constructing any prediction method. Extracting a training set for testing of prediction methods is largely from database annotations and this can involve a lot of work retrieving huge data from databases for training/testing. The resultant effect, more often than not, is the introduction of error into the desired goal.

One problem that is rarely properly addressed, especially when training global property methods, is homology in the data. All machine-learning methods have many free parameters and therefore have the potential to overfit such that they learn each example “by heart” instead of learning the general pattern of the training data, including noise contained in the data. An overfitted machine-learning method that is able to reproduce all its training patterns exactly will typically have a bad generalization ability resulting in low predictive performance. Ideally, all training examples should have experimental evidence and not be inferred by similarity or existing prediction methods. (Emanuelsson *et al*., 2007). But we do also know that strictly adhering to the use of data that have experimental evidence is not possible at all situations. For example it might be impossible to ever train a method using experimentally proven sequences when working with rare or poorly characterized sorting signals. The advantage, however, is that machine learning

methods are potentially designed in way that they are able to deal reasonably with noisy data. Neural networks, for example have been seen to learn erroneous examples at a dragging rate or even refuse to learn them at all in the first place. The ability to diagnose noisy data as well as the ability to decide not to learn at all in neural networks is strongly determined by the construction of the model, in terms of learning procedure and model complexity.

Proteins entering the non-classical secretion pathway cannot be correctly identified using prediction methods such as PSORT (Nakai and Horton, 1999) and SignalP (Bendtson, 2004). The reason for this might be that they that utilize alternative mechanisms or they have "hidden" signals.

The aim of this research was to test the hypothesis that this "missing link" between non-classically secreted proteins and existing secretory signal could be resolved by generating alternative gene predictions for these proteins It is also hypothesized that these proteins could have hidden signals. To do this, some of the existing methods for secretory signal predictions were identified for the tasks. The methods used in this work link a set of existing secretory signal prediction tools and tries to interpret and compare the outputs from all the methods. The difference between the approach here and all other secretory signal prediction tools is that it takes the prediction one step backwards by starting from the DNA level. The method used here combines the very well know secretory signal prediction methods. It combines SignalP 3.0, Phobius, PSORT as well as LocateP. To further give a sense of reliability to the outputs of the methods just enumerated, this work also utilizes tools to screen for ribosomal binding sites as well as conservation analysis with special interest around the protein start sites

By starting predictions from the DNA level, it is believed that the start site will be hugely optimized. Many of the annotations of proteins in databases are largely prone to error. These errors are largely due to annotating proteins with wrong start sites. It is common knowledge that wrong annotation leading to wrong start sites will inherently lead to a wrong signal prediction. The idea of utilizing the cDNA of the protein is believed to improve prediction accuracy. To optimize the cDNA of the target protein, the flanks are

extended hundreds of base pairs (in this work, 500 basepairs was used) and the resulting cDNA read in all six frames, split into all possible open reading frames, before being translated back to the protein. The rationale behind extending the DNA flanks is to see how well the annotated or "known" start site was correctly predicted. It is also a useful way of knowing if some other "qualified" codon could have been the start site. In other words, extending the flanks gives an idea of the integrity of the annotated start site, at least as far as this work is concerned.

During translation, the bacterial codon table was used such that codons differing from "ATG" were also translated as potential start sites. Start site optimization and splitting of the DNA/protein into all possible orfs was achieved by a perl script as well as manual inspection. Each of the orfs from above was analyzed using the secretory signal tools stated above.

The choice of tools used in this research has been influenced, to an extent, by the intented tasks and expected results. SignalP, for example, uses a combination of NN and HMM models in its predictions. It is known that neural networks based tools can minimize errors in their predictions by reducing the effects of noisy data. SignalP is also one of the most frequently used signal peptide prediction tools. Even at that it has its own shortcomings, sometimes returning false positive predictions and in some cases false negative prediction as well. Phobius, on the other hand has the advantage of being able to predict the presence of a signal peptide as well as a transmembrane topology. However, Phobius, more often than not leaves us with an extra task of determining if its positive predictions are actually true. In other words, a prediction made to be "non-cytoplasmic" by Phobius, for example, does not necessarily imply that there is a signal peptide present and vice versa. Phobius appears to be superior to most other methods when it comes to prediction of transmembrane helices close to the N terminus, which are easily confused with signal peptides (Bendtsen, 2004). PSORT was originally designed for Gram-positive bacteria, and this is particular in its choice here as all the proteins used in this research are from Gram-positive bacteria.

It is clear from the above that no single method of secretory signal prediction is all encompassing. Current research, such as this one, and elsewhere, has utilized a combination of tools with a view to enhancing prediction accuracy. Before now, three

prominent methods, TargetP, SignalP, and SignalPeptidePrediction have been used to analyze the precursor sequences of human cyclooxygenases 1 and 2(COX-1, COX-2). The above provides an example of how more than one prediction tool can be combined to enhance accuracy. The three different prediction tools produce consensus potential SP I cleavage sites (Schneider and Fechner, 2004). Here we have combined SignalP, Phobius, PSORT, and LocateP primarily for the prediction of signal peptides and or cleavage sites. Additionally Patser v.3e (van Helden, 2003) and entropy calculating tool (Capra and Singh, 2007) were used jointly alongside the signal peptide prediction tools. The goal is to be able to compare outputs from the individual tools and determine significance/accuracy of predictions or otherwise. Homology searches and subsequent alignment of multiple sequences gives an idea of how conserved the target protein is when compared with its homologues. The idea is that same proteins should have very similar composition, at least, in terms of primary structure, and more like to have the same start site.

It is believed that a combination of tools will lead to a better prediction than when the tools are used individually. Since I was trying to access the possibility of obtaining signal peptide prediction by alternative gene predictions, there was the need to look at sites where the ribosome would probably bind and compare the results, first, with predictions by other tools as well as experimentally known facts where it exist. Since ribosomes are basically required to be present at specific sites for translation initiation.

6.2 Alternate gene predictions may enhance signal peptide detection in non-classical secretion

Results from the analysis of data in this work points to an improvement in signal peptide prediction using the methods presented here. 305 sequences generated from the optimization of the original sequences were analyzed. Out of these, 53 were from arginine, 80 from glutamine synthase while 58, 64, and 50 were from EF-Tu, GroEL, and enolase respectively. When each of these sequences was analyzed by SignalP 3.0, the

result did not show only an appreciable improvement in the prediction based on the value of the D-score, there was actually a signal peptide prediction in orf 59 of GroEL. The values from the best D-scores for both the original protein and the optimized orfs are shown in Table 5.1. A comparison of the signal peptide prediction in. the original and optimized sequences with respect to the D-score cut-off is shown in Figure 5. 2. Even though there was no actual signal peptide prediction for enolase, a score of 0. 411 versus the D-score cut-off of 0.45 is to my mind significant enough to believe that there is a possibility of finding a signal peptide at orf 10 of enolase. This could be justified by the actual prediction of a signal peptide by Phobius in the same orf in enolase. For arginase, glutamine synthase, and EF-Tu an increase by 153%, 226%, 114% in the value of best D-scores strongly suggests that the method used in this research is very promising. As the goal of this work was to implement a set of signal peptide prediction tools, it is only useful to examine the results of the other tool before drawing conclusions. We note however that SignalP 3.0 HMM results are not described here since they were largely insignificant.

Results from Phobius tended to support the predictions by SignalP 3.0. Phobius also predicted the presence of signal peptides in the same orf as in enolase and GroEL. This time there was an actual prediction for enolase. It is interesting to note that Phobius also predicted a transmembrane topology in orf 9 of enolase fueling the suspicion that the orfs 9 and 10 in question may in fact contain a signal peptide.

Majority of the predictions by PSORT were observed to be predicted as “cytoplasmic”. The orfs in which “extracellular” predictions were made have been shown in Table 5.4.PSORT predicted the orf 60 in GroEL and orf 1 in enolase to be extracellular. This result also supports to some extent the prediction trend already seen for both proteins from SignalP 3.0 and Phobius analysis. LocateP prediction results were from the original original protein sequences only and they serve for comparisons only.

The motivation for this research was to try and find why some known proteins are secreted even though existing signals don’t predict signal peptides in their sequences. The hypothesis presented here is that the proteins could have wrong translation start sites or

they might be using a novel signal. In order to minimize wrong start site annotation, the original proteins were first optimized by exploring all possible start sites and checking for signal peptides in all of these. The results from SignalP 3.0, Phobius and PSORT show significant signal peptide detection from the optimized protein sequences compared with the original sequences. To further check the quality of the method applied here, the results from signal peptide analysis tools were compared with homology and RBS analysis. As shown in Tables 5.6 and 5.7, and Figure 5.3, it is easy to see how conserved the region around the start site is for the original sequences especially enolase and GroEL. Conservation as reported earlier was calculated as the average of +/-10 residues on both sides of the start site of the original protein using scores returned from the conservation calculating tool. Depending on the RBS matrix used (here we have used the two popular ones, AGGAGG and AAGGAG) the positions of the RBSs offer useful guide in trying to arrive at correct prediction. In the case of enolase, for example, it is easy to see that if the ribosome binds at position 501, then the prediction at the orf 10 is likely possible. Further, a small entropy score indicates a better conservation. In the case of enolase, a conservation score of 0.7406 would suggest a weak conservation, but more importantly a combination of SignalP, phobius, and PSORT points to a positive prediction.

GroEL, on the other hand, is a little bit trickier. Two possible RBSs were found for GroEL at the DNA positions 142 and 500 respectively, but a signal peptide was detected in GroEL's orf 59. The connections between these parameters are not very remote since orf 59 is located towards the end of the DNA, yet they offer very interesting insights into many possibilities. In this case, it is believed an alternative prediction is able to provide the signal peptide used by the protein for secretion. A conservation score of 0.9164 for GroEL implies a low, if not poor, conservation around the start site of original protein and thus supports the prediction of a signal peptide at orf position of GroEL above.

The predictions by a combination of all the tools used in this work did not yield any signal peptide predictions for arginase, EF-Tu, and glutamine synthase. However, there were significantly appreciable indications that the predictions obtained from the

optimized sets of proteins sequence were better than that of the original sequence especially those from SignalP.

6.3 Future experimental approaches

Most of the work done in this work has utilized individual tools to analyze data. Most of the data preprocessing was done manually. This obviously limits one to the amount of data that can be analyzed. As an approach to this work in the future, the individual tools could be linked to make one tool. This will help reduce possible errors from feeding data to each tool one at a time. Another advantage from linking the tool is that a very large amount of data can be analyzed with speed, and possible errors eliminated.

7. Conclusion

The aim of this research was to find signal peptides in known non-classically secreted proteins by screening all proteins sequences generated by alternative predictions relative to the original “non-classical proteins”. Some of the best known methods for signal peptide prediction have been utilized here to analyze these sequences. The combined predictions from these methods have been analyzed and the results have shown that alternative gene prediction methods may be useful in searching for signal peptides.

The prediction of signal peptides could be enhanced by first optimizing the start site of the starting protein. This can be achieved by carrying out alternative gene predictions for the protein. At the same time, the methodology used in this work can assist in the design of signal peptide prediction tools.

References

Antelmann, H., Tjalsma, H., Voigt, B., Ohlmeier, S., Bron, S., van Dijl, M. and Hecker, M.A. (2001). A proteomic view on genome-based signal peptide predictions. *Genome Res*, **11**, 1484- 1520.

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

Bendtsen, J.D., Kiemer, L., Fausboll, A. and Brunak, S. (2005). *BMC Microbiology*, **5**, 58.

Bendtsen, J.D., Nielsen, H., Von Heijne, G. And Brunak, S. (2004). Improved prediction of Signal peptides :SiganalP 3.0. *J. Mol. Biol*., **340**, 783-795.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003). The SWISS-PROT knowledge database and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 361-370

Byvatov, E. and Schneider, G. (2003). Support vector machine applications in bioinformatics. *Appl. Bioinformatics*, **2**, 66-67.

Capra, J.A. and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875-1882.

Capra, J.A. and Singh. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875 – 1882.

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

Chou, K.C. and Elrod, D.W. (1999). Protein of subcellular location Prediction. *Protein Eng*., **12**, 107-118.

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

Chou, K.C. and Shen, H.B. (2006). Predicting eukaryotic protein sub cellular localization by fusing optimized evidence-theoretic k- nearest neighbour classifiers. *J. Proteome Res*., **5**, 1888-1897.

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

Constanzo, L.D., Moulin, M., Haertlein, M., Meilleur, F. And Christianson, D.W. (2007). Expression, Purification, Assay and Crystal Structure of perdeuterated Human rginase I. *Arch Biochem Biophys.*, **465**, 82-89.

Cristianini, N. and Shawe-taylor, J. (2000). An introduction to Support Vector Machines and Other Kernel-based Learning methods. Cambridge Press, Cambridge.

Darnell, J., Lodish, H. And Balyimore, D. (1990). Molecular Cell Biology. Freeman, Newyork.

Dilks, K., Rose, R.W., Hartmann, E. and Pohlsschroder. (2003). Prokaryotic utilization of the twin-arginine translocation pathway: a genomic survey. *J Bacteriol*, **185**, 1478-1483.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001). Pattern Classification. John Wiley &Sons, New York.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acid Res*, **32**, 1792-1797.

Emmanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocol*, **2**, 953-971.

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

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M. And Brinkman, F.S.L. (2005). PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*, **21**, 617-623.

Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K. and Brinkman. (2003). PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acid Res*, **31**, 3613-3617.

Harth, G. and Horwitz, M.A. (1997). Expression and efficient export of enzymatically active *Mycobacterium tuberculosis* glutamine synthetase in Mycobacterium smegmatis and evidence that that the information for export is contained within the protein. *J Biol Chem*, **272**, 22728-22735.

High, S. and Dobberstein , B. (1992). Mechanisms that determine the transmembrane disposition of proteins. *Curr. Opin. Cell Biol.* , **4**, 581–586.

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

Hirose, I., Sano, K., Shioda, I., Kumano, M., Nakamura, K. and Yamane, K. (2000). Proteome analysis of Bacilluus subtilis extracellular proteins: a two-dimensional protein electrophoresis study. *Microbiol*, **146**, 65- 75.

Horton, P., Park, K-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res*., **35**, W585-W587.

Jefffery, C.J. (2003). Moonlighting proteins: old proteins learning new tricks. *Trend Genet*, **19**, 415-41.

Juncker, A. S., Willenbrock, H., von Heijne, G., Brunak, S., Nielson, H. And Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*., **12**, 1652-1662.

Käll, L., Krogh, A. And Sonnhammer, E.L.L. (2004). A cobined Transmembrane Topology and Signal peptide Prediction Method. *J. Mol. Bio*., **338**, 1027-1036.

Käll, L., Krogh, A. And Sonnhammer, E.L.L. (2007). Advantages of Combined transmembrane topology and signal peptide prediction – the Phobius web server. *Nucleic acids Res*, **35**, W429-W432.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern*., **43**, 59-69.

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

Lee, P.A., Tullman-Ercek, D. And Georgiou, G. (2006). The bacterial twin-arginine translocation pathway. *Annu Rev Microbiol*, **60**, 373-395.

Lenz, L.L., Mohammadi, S., Geissler, A. And Portnoy, D.A. (2003). SecA2-depent secretion of autolytic enzymes promotes *Listeria monocytogenes* pathogenesis. *Proc Natl Acad Sci*, **100**, 12432-12437.

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

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

Martogolio, B. And Dobberstein, B. (1998). Signal sequences: more than just greasy peptides. *Trends in cell Biol*, **8**, 410-415.

Menne, K., Hermjakob, H. And Apweiler, R. (2000). A comparison of signal sequences prediction methods using a test set of signal peptides. *Bioinformatics*, **16,** 741-742.

Miethke, M., Klotz, O., Linne, U., May, J.J., Beckering, C.L. and Marahiel, M.A. (2006). Ferri-bacillibactin uptake and hydrolysis in bacillus subtilis. *Mol Microbiol*, **61**, 1413-1427.

Nair, R. and Rost, B. (2008). Predicting protein subcellular localization using intelligent systems. *Methods Mol. Biol*., **484**, 435 – 463.

Nakai, K. and Kanehisa, M. (1991). Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins: Struct. Funct. Genet.*, **11**, 95-110.

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

Nelson, D. L. and Cox, M. M. (2005). Lehninger Principles of Biochemistry. 4th Edition, W. H. Freeman and Company, New York.

Nicholas, K.B., Nicholas H.B. Jr. and Deerfield, D.W. II. (1997) GeneDoc: Analysis and Visualization of Genetic Variation, *EMBNEW.NEWS*, **4,** 14.

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

Nielson, 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: Struct. Funct. Genet*., **26**, 165-177.

Nielson, H., Engelbtecht, J., Brunak, S. And von Heijne, G. (1997). A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of cleavage sites. *Int. J. Neural Syst*., **8**, 581-599.

Ollinger, J., Song, K.B., Antelmann, H., Hecker, M., and Helmann, J.D. (2006). Role of the Fur regulon in iron transport in Bacillus subtilis. *J Bacteriol*, **188**, 3664-3673.

Palmer, T. AND Berks, B.C. (2003). Moving folded proteins across the bacterial cell membrane. *Microbiol*, **149**, 547-556.

Park, K. –J. and Kanehisa, M. (2003). Prediction of subcellular location by support vector machines using composition of amino acids and amino acid pairs. *Bioinform*, **19**, 1656-1663.

Pugsley, A.P. (1993). The complete general secretory pathway in Gram-negative bacteria. *Microbial. Rev*., **57**, 50-108.

Reinhardt, A. And Hubbard, T. (1998). Using neural networks for the prediction of the subcellular location of proteins. *Nucleic Acids Res*., **26**, 2230-2236.

Ren, Q.,Kang, K.H. and Paulson, I.T. (2004). TransportDB: a relational database of cellular membrane transport systems. *Nucleic Acids Res*, **32**, D284- D288.

Rubartelli, A., Bajetto, A., Allavena, G., Wollman, E., and Sitia, R. (1992). Secretion of thioredoxin by normal and neoplastic cells through a leader-less secretory pathway. *J Biol Chem*, **267**, 24161-24164.

Sadowski, J. and Kubinyi, H. (1998). A scoring scheme for discriminating between drugs and non-drugs. *J. Med. Chem*., **18**, 3325-3329.

Sargent, F., berks, B.C. and Palmer, T. (2006). Pathfinders and trailblazers: a prokaryotic targeting system for transport of folded proteins. *FEMS Microbiol*, **49**, 1377-1390.

Schaumburg, J., Diekmann, O., Hagendorff, P., Bergmann, S., Rohde, M., Hammerschmidt, S., Jansch, L., Wehland, J. and Karst, U. (2004). The cell wall proteome of *Listeria monocytogenes*. *Proteomics*, **4**, 2991-3006.

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

Schneider, G. (1999). How many potentially secreted proteins are contained in a genome? *Gene*, **237**, 113-121.

Schneider, G. and So, S.S. (2003). Adaptive systems in Drug Design. Landes Bioscience, Georgetown, TX, **pp**. 87-88.

Schneider, G. and Wrede, P. (1998). Artificial Neural Networks for computer-based molecular designs. *Prog. Biophys. Mol. Biol*., **70**, 175-222.

Schneider, P., Tanrikul, Y. and Schneider, G. (2009). Self-organizing maps in drug discovery: compound library design, scaffold-hopping, repurposing. *Curr Med Chem*., **16**, 258-266.

Scott, M. Thomas, D. And Hallet, M. (2004). Predicting subcellular localization via protein motif co-occurence. *Genome Res*., **14**, 1957-1966.

Tjalsma, H. and van Dijil, M. (2005). Proteomics-based consensus prediction of protein retention in a bacterial membrane. *Proteomics*, **5**, 4472- 4482.

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

Van Helden, J. (2003). Regulatory Sequence Analysis Tools. *Nucleic Acids Res*, **31**, 3593-3596.

Vanet, A. and Labigne, A. (1998). Evidence for specific secretion rather than autolysis in the release of some *Helicobacter pylori* proteins. *Infect Immun*, **66**, 1023-1027.

Villa, E., Sengupta, J., Trabuco, L.G., LeBarron, J., Baxter, W.T., Shaikh, T.R., Grassucci, R.A., et al., (2009). Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis. *PNAS*, **106**, 1063-1068.

Vitikainen, M., Lappalainen, I., Seppala., R., Antelmann, H., Boer, H., Taira, S., Savilahti, H., Hecker, M., Vihinen, M., Sarvas, M. And Kontinen, V.P. (2004). Structure-function analysis of PrsA reveals roles for the parvulin-like and flanking N- and C-terminal domains in protein folding and secretion in *Bacillus subtilis*. *J Biol Chem*, **279**, 19302-19314.

Von Heijne, G. (1986). A new method for predicting signal cleavage sites. *Nucleic Acids* Res., **14**, 4683-4690.

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

Yu, C., Chen, Y., Lu, C. and Hwang, J. (2006). Prediction of Protein Subcellular Localization. *Proteins: Structure, Function, and Bioinformatics*, 64, 643-651.

Yu, C.S., Lin, C.J. and Hwang, J.K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci*, **13**, 1402-1406.

Yu,C.S., Chen,Y.C., Lu, C.H. and Hwang, J.K. (2006). Prediction of subcelluar via protein motif co-occurrence. *Proteins*, **64**, 643-651.

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

Zhou, M., Boehhorst, J., Francke, C., and Siezen, R.J. (2008). LocateP: Genome –scale subcellular-localization predictor for bacterial proteins. *BMC Bioinformatics*, **9**, 173.