Chapter 3

Lokum: *ab initio* protein subcellular localisation prediction for eukaryotes by using mono and bipartite motifs, transmembrane protein topologies, and amino acid composition

3.1 Introduction

Protein sorting in eukaryotes is generally more complicated than in bacteria, simply because a typical eukaryotic cell contains a larger number of compartments. Presence of different compartments defined by various internal membranes within the cell mean different proteins must successfully pass through these internal envelopes, which naturally involves a larger number of molecules and different targeting and retention mechanisms. Identification of protein regions that are involved in protein transport across a certain membrane is a key step in all prediction efforts mimicking the underlying biological interactions. I try to address this issue by using a new, probabilistic, *ab initio* protein motif discovery tool,

NestedMICA (Down & Hubbard, 2005), which has been recently shown to work better than another popular program MEME (Bailey & Elkan, 1995), particularly for short proteins motifs that range in 3-9 amino acids (aa) (see Chapter 2 or Doğruel *et al.* (2008)). This makes NestedMICA suitable for use in localisation signal discovery, as targeting signals could be as short as 3 aa. NestedMICA, using a new Monte Carlo technique called Nested Sampling (Skilling, 2004), reports motifs in Position Weight Matrices (PWMs).

One of the basic forms protein localisation signals could be characterised by are multi-component probabilistic motifs, which most motif finders cannot deal with. I use a combinatorial strategy involving both NestedMICA and the Eponine tool (Down & Hubbard, 2002) that I have improved for protein sequence support.

3.1.1 Features used in Lokum

In this study, to predict protein localisation I used mono- and bipartite protein localisation signals discovered by NestedMICA and Eponine, other NestedMICA motifs that are not directly involved in localisation but that I show to be useful in the computational predictions, amino acid frequency distributions, and finally protein transmembrane topology statistics.

Apart from the difficulty of discovering genuine localisation signals, in signal-based *ab initio* protein subcellular localisation prediction another complication is the poor discriminative power of these motifs in the classification problem. Proteins can share the same type of localisation motifs, not necessarily because they are from the same cellular localisation, but because they could be involved in a similar translocation pathway. Partly because of such common localisation

signals, it is usually difficult to attain high prediction accuracies in automatic *ab initio* classification methodologies (for a list of some popular automatic prediction tools, see Section 1.1.1 in the introduction chapter). One possible way to reduce the weaknesses of individual features is to use as many relevant protein properties in combination as we can, where a pre-trained automatic prediction system will evaluate possible relations among the features to make a final decision. I used a popular classification method, Support Vector Machines (SVM), as they can provide very good generalisation performance by finding optimal hyper-surfaces that split data points of different classes in multi-dimensional spaces.

One general type of intrinsic signals proteins carry is targeting sequences. They are usually found in the N-terminal regions of proteins, and some of them are cleaved off from the nascent protein after the protein is translocated across a membrane. There could also be targeting signals located on the far C-terminus, like the Peroxisomal Targeting Signal 1 (PTS1) which is usually characterised by the tripeptide sequence SKL (Gould *et al.*, 1987, 1989). However, PTS1 is not found in all proteins that are post-translationally transported to the peroxisome. It is believed that peroxisomal proteins contain a weakly conserved N-terminal signal of the form [RK][LVI].....[HQ][LA], named PTS2 for "Peroxisomal Targeting Signal 2", where the dots represent any amino acid (Osumi *et al.*, 1991; Swinkels *et al.*, 1991). Certain mitochondrial targeting peptides are located in the N-terminus, too, while these proteins can also have secondary signals which are thought to be present possibly anywhere along the entire pre-protein sequence (Endres *et al.*, 1999; Wiedemann *et al.*, 2001).

Not all secreted proteins have N-terminal targeting signals (Bendtsen *et al.*,

2004a; Nickel, 2003), however the major type of proteins that have an N-terminal targeting signal is the secretion pathway proteins, as they contain a conserved signal peptide (SP) (Milstein *et al.*, 1972) that can range in length between 20 and 30 amino acids in eukaryotes (Emanuelsson *et al.*, 1999; von Heijne, 1990). A usually cleavable N-terminal targeting peptide directs them into the ER by penetrating through the ER membrane, while the rest of the nascent polychain peptide is still being synthesised in ribosomes that are located near the ER. A smaller number of them are maintained and employed by the ER if they contain the tetrapeptide KDEL signal on their C-termini (Pelham, 1995). Most of these proteins that pass several "quality control tests" of the ER are then sent to the Golgi apparatus for further processing, but some of them, such as the malfolded or unassembled ones that failed those tests, are delivered by the ER to the proteolytic system for degradation. This indicates there is some sort of back-and-forth traffic between the ER and Golgi, but that there are no reported retention or targeting motifs associated with the Golgi compartment. However, Yuan & Teasdale (2002) showed that up to a certain extent it is possible to distinguish Golgi Type II membrane proteins from the others, by using the hydrophobicity values and frequencies of different residues within their transmembrane domains. For most cargo molecules traversing through the "Golgi cisternae", or multiple ordered stacks of the Golgi apparatus, Golgi acts only as an intermediate place. They eventually either end up in the plasma membrane, or are secreted out of the cell.

N-linked glycosylation is a common type of post-translational protein modification that takes place shortly after the nascent chain enters into the ER lumen

(Kaplan *et al.*, 1987; Machamer *et al.*, 1985). Starting as early as 1985, some previous studies have claimed glycosylation could have a role in cell transport (Guan *et al.*, 1985; Hannink & Donoghue, 1986; Kelley & Kinsella, 2003; Yan *et al.*, 2002) while others (Matsuda *et al.*, 2004; Mohrmann *et al.*, 2005) reported that it is not specifically required in cell surface transport for the tested protein molecules. A recent study demonstrated that N-linked glycosylation is required for structural stabilisation but not for membrane localisation of a tested particular protein (Gao & Mehta, 2007). The generally accepted notion seems to be that N-linked glycosylation is not directly involved in localisation. However, I show in this chapter that it is enriched in secretory pathway proteins over the other types, making it a potential secondary signal to aid in computational localisation prediction, just in a similar way to use the "secondary signal" coming from protein composition.

Amino acid residues can have similar physical and chemical characteristics. It is for this reason that protein signals such as the secretory pathway signal peptide (SP) are described often in terms of their general characteristics like hydrophobicity, net charge etc., rather than in terms of their individual amino acid letters which might not be conserved, as in the case of SP, for example. Individuals of different generations can have protein sequences that are still functionally similar yet different in terms of the actual amino acid line up due to the associated DNA-level mutations that take place in the process of evolution. Up to a certain extent, it is therefore possible to safely substitute certain amino acid residues with other similar ones without much harming the function and thus affecting the tertiary structure of a protein. The study of such functionally homologous blocks

showing sequence variation has resulted in amino acid substitution matrices like PAM (Eck & Dayhoff, 1966) and BLOSUM (Henikoff & Henikoff, 1992)

In Lokum, apart from the motifs discovered *ab initio*, I used the normalised amino acid abundance rates in sequences. Amino acid composition has been proven useful in localisation prediction (Klein *et al.*, 1984; Nakai & Kanehisa, 1991; Reinhardt & Hubbard, 1998). There have been many machine learning approaches incorporating amino acid frequency distributions alone or sometimes accompanied with other features. Reinhardt & Hubbard (1998) suggested that using amino acid composition would be advantageous over other signal-based methods as it makes a protein less susceptible to possible annotation errors, particularly in the 5' regions where most targeting signals reside. However, by using probabilistic representations such as Position Weight Matrices (PWMs) to characterise such signals it is possible to tolerate slight sequential variations. This argument becomes more valid especially for PWM positions having almost flat distributions of amino acid probabilities, where any amino acid can be expected to occupy those positions.

The third type of protein feature I used is predicted secondary transmembrane structures. Amongst the transmembrane topology predicting programs such as TopPred (Claros & von Heijne, 1994), SOSUI (Hirokawa *et al.*, 1998), TMHMM (Krogh *et al.*, 2001) and HMMTOP (Tusnády & Simon, 2001), studies on evaluation of these programs showed that TMHMM performed better than the rest of the predictors. It has been reported that, in general all the tested programs can easily misclassify the predominantly hydrophobic membrane spanning regions as N-terminal signal peptides which also contain a similar strong hydrophobic re-

gion (Lao *et al.*, 2002; Müller *et al.*, 2001). For the same reason, signal peptide predictors may often misjudge transmembrane regions as signal peptides, too. Chapter 5 summarises my efforts to develop a transmembrane topology predictor that can be used in subcellular localisation prediction, but this HMM-based tool didn't perform as well as TMHMM. Therefore, in the end, in Lokum I used transmembrane topology statistics based on TMHMM predictions.

3.1.2 Predicted classes

In this manuscript, I compare my *ab initio* method, Lokum (Localisation prediction using motifs) with both PSORT and MultiLoc. Similarly to these programs, Lokum predicts nine localisation categories for animal proteins: nucleus, cytoplasm, plasma membrane, extracellular space, mitochondrion, endoplasmic reticulum, Golgi apparatus, lysosome, and peroxisome. Next, substituting lysosomes with vacuolar proteins in the animal set, Lokum's predictions are extended to cover all major nine fungal protein localisations, and finally ten plant localisation classes with the addition of chloroplast to the list of fungal classes.

3.2 Materials and methods

3.2.1 Localisation motif discovery with NestedMICA

I used NestedMICA, an *ab initio* DNA and protein motif discovery tool, to search for localisation-specific motifs that can be used in classification. NestedMICA employs a new Monte Carlo inference technique called nested sampling developed by Skilling (see page 28). It was originally developed for finding DNA motifs, and has been recently extended to find protein motifs (see Chapter 2 or Doğruel *et al.*

(2008)). It reports motifs in the form of Position Weight Matrices (PWMs) which allow more flexibility for having alternative residues at certain positions than, for example, motifs represented as regular expressions.

The target motif length interval parameter was given to be between 3 and 15 amino acids long in all the NestedMICA runs. Initially, the target motif number was specified as 2, but I experimented with this program parameter for each localisation class to cover as many potentially localisation-related motifs as possible. NestedMICA was run on the full-length sequences, as well as 50 N- and C-terminal amino acid chunks for each localisation dataset. The ER retention signals (Figure 3.4f-g) and PTS1 (Figure 3.4j) were recovered when NestedMICA was fed with the last (C-terminal) 50aa long regions.

For motif discovery purposes, I used nine datasets from pTarget (Guda & Subramaniam, 2005), a subcellular localisation predictor based on searching more than 2100 PFAM domains, after reducing the mutual sequence identities of the datasets from 95% to a maximum of 40% by the CD-HIT algorithm (Li & Godzik, 2006). Table 3.1 lists the number of sequences before and after applying redundancy reduction. For localisation categories that do not exist in pTarget, namely for chloroplasts and vacuolar classes, I used the redundancy-reduced datasets of MultiLoc (Höglund *et al.*, 2006), a recent subcellular localisation prediction program. The details for these two sequence sets can be seen in Table 3.4 where Lokum predictions are compared with those of MultiLoc. I further decreased the maximum mutual sequence identities of the MultiLoc datasets as well to 40% before running NestedMICA.

NestedMICA uses complex background models which could be composed of

Localisation class	Number of sequences in the original set	Number of sequences after the filtering
Cytoplasmic	2062	946
ER	693	251
Extracellular	5688	1671
Golgi	221	141
Lysosome	174	66
Mitochondria	1698	711
Nuclear	3446	2014
Peroxisome	173	83
Plasma membrane	4162	1212

Table 3.1: **Sequences used in the motif discovery phase.** Each pTarget (Guda & Subramaniam, 2005) dataset, originally having a sequence identity of 95%, was filtered to have a maximum mutual identity of 40% by using the CD-HIT (Li & Godzik, 2006) clustering program. Vacuolar and chloroplast classes do not exist in pTarget, so the corresponding datasets of MultiLoc (Höglund *et al.*, 2006) were used for these two categories (Table 3.4).

multiple subgroups of different amino acid probability distributions to better represent different sequence regions statistically inclined to feature certain amino acid residues more frequently. As has been discussed in 2.3.1, training dedicated background models for each sequence dataset yields better performance than using a generic background model. Therefore, for each type of localisation a specialised NestedMICA background model was trained. The background model parameters used for each localisation dataset are summarised in Table 3.2. NestedMICA was run on each dataset with its default protein motif finding parameters.

3.2.2 Motif selection

NestedMICA does not report any significance measure. To decide if a reported motif is significantly contributing to localisation classification, I scanned it across

some test sequences to plot Receiver Operating Characteristic (ROC) curves, as in Figures 3.7 and 3.9. A motif discovered from the plasma membrane set, for example, was tested for its usefulness to discriminate between plasma membrane sequences and every other class of sequences. By using equal number of sequences of both types, in each binary classification based on only raw bit scores of a motif, I classified sequences in two classes according to a range of motif score thresholds. ROC curves were plotted using the sensitivity and specificity pairs obtained for each threshold used. Motifs producing promising ROC curves in any possible binary classification were then selected to be used in the general multi-class SVM.

Additionally, I performed a brute-force principle component analysis to assess the contribution of each selected feature, or dimension of SVM vectors. I observed

Dataset	MC-order	Number of Mosaics
ER	0	5
Vacuolar	0	2
Lysosome	0	4
Golgi	0	5
Mitochondria	1	3
Chloroplast	0	5
Peroxisomal	0	6
Nuclear	1	4
Cytoplasmic	1	6
Extracellular	1	4
Plasma membrane	1	6

Table 3.2: **Protein background parameters for datasets used in localisation motif discovery** The table summarises the NestedMICA background properties of the datasets where localisation related motifs were searched, in terms of the used Markov-chain order and the number of mosaic classes in the background. These parameter values have been optimised after a systematic analysis of each dataset as described in Chapter 2, Section 2.2.4.

the effects of removing a single or multiple dimensions from the input vectors on the overall performance. Features increasing the prediction performance upon removal were not used in the final SVM. None of the amino acid frequency dimensions were necessary to remove. As an interesting example, PTS2 was among the motifs I decided not to use in the end (see results).

3.2.3 Using Eponine with NestedMICA for multi-component motif discovery

Some localisation signals can consist of multiple components separated by a distance. The best known such signal is the bipartite NLS (Dingwall & Laskey, 1991) which has been identified to have two core NLS parts that are separated by at least 10 (Robbins *et al.*, 1991) and around 12 (Schreiber *et al.*, 1992) "spacer" amino acids. NestedMICA currently does not deal with multi-component motifs. I modified and extended the Eponine (Down & Hubbard, 2002) tool to discover and represent such protein localisation signals.

Eponine was originally developed to find promoter models from mammalian genomic DNA to represent multi-component, hiearchical motifs. Eponine describes these multi-component motifs as Eponine Anchored Sequence (EAS) models, where motifs are modeled around a fixed, or "anchor" point. It generates a number of weight matrices corresponding to different sequence motifs which it believes to be collectively involved in signaling a certain sequence characteristics. Each motif within an Eponine motif set has a positional distribution relative to a point of interest, such as a transcription start site (TSS) point. When scoring sequences with an Eponine model, positional deviations of the best matching

sub-motifs with respect to the means of the corresponding Gaussian distributions are considered, too. Figure 3.1 shows an EAS which models mammalian TSS regions, as reported by Down & Hubbard. It has been later shown in a PhD dissertation that it can actually be used as a multi-purpose motif finder, where it was specifically used in the detection of transcription termination sites (TTS) (Ramadass, 2005). Figure 3.2 shows the discovered EAS model for mammalian TTS regions.

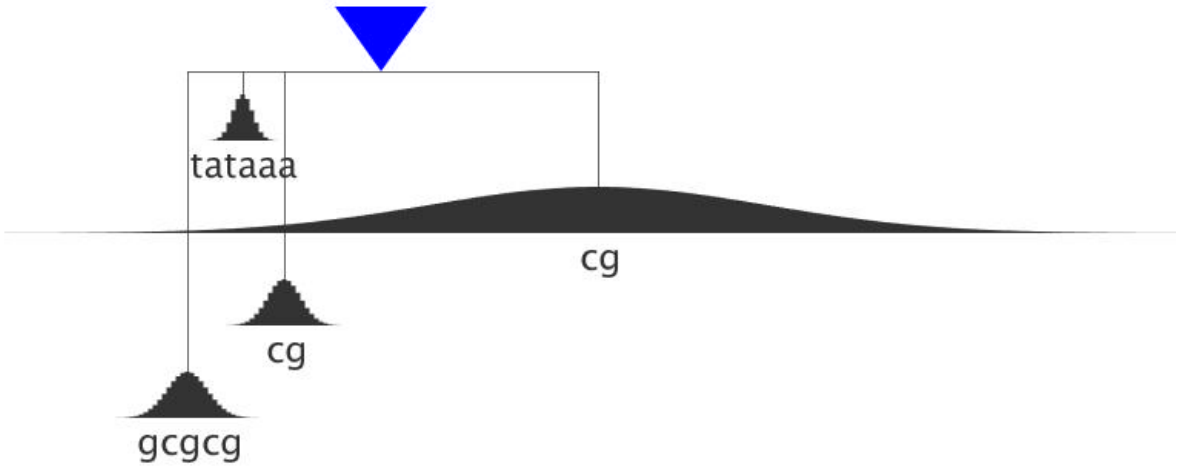


Figure 3.1: **Eponine TSS model.** Blue triangle in Eponine Anchored Sequence (EAS) models indicate the anchor point. Individual motif weight matrices are positioned with respect to the anchor. Gaussian distributions indicate the positional distributions of the corresponding motif. This TSS model has been reproduced from the original Eponine publication (Down & Hubbard, 2002).

Eponine was later extended for use in non-coding DNA region analysis with the purpose of discovering overrepresented multi-component motifs conserved in mouse and human intergenic regions Down & Hubbard (2004). This version of Eponine describes motifs as Eponine Windowed Sequence (EWS) models, in analogy to the previous model type. In EWS, unlike the first version, there

Figure 3.2: **Eponine TTS model.** As in Figure 3.1, this mammalian transcription termination site (TTS) model, too, is an example to EAS models, as reported in Ramadass's PhD thesis.

is no need to have a specific sequence position around which other sub motif components are placed. Instead, this model classifies sequence regions based on only their sequence contents within specific windows.

Eponine, which uses Biojava (BioJava, 2007) libraries, relies on a new machine learning strategy called Relevance Vector Machines (RVMs) (Tipping, 2001) taking a set of suggested basis functions and then iteratively choosing certain combinations that would presumably yield a better performance at each step. To this end, it optimises candidate PWMs and their parameters including width and positional Gaussian distributions. It requires both a positive and a negative training set to decide if the combination used at each step is better discriminating the two classes. Because Eponine actually works by trying to discriminate data points, it searches for motifs in the negative set, too. This can result in reported models to have some "negative" motifs which have negative weights in the models (they are drawn in blue colour in the graphical representations, as opposed to the black "positive" motifs). Generally speaking, not having any negative motifs in reported Eponine models trained using negative datasets that are obtained by

shuffling the used positive samples indicates a successful training. More information about how Eponine works can be found in the original Eponine publication by Down & Hubbard (2002).

With the idea of employing Eponine to discover multi-component protein motifs such as the bipartite nuclear motifs, I added protein sequence support to Eponine. Having modified it to accept protein sequence input, I tested its efficiency in the protein space. However, my tests generally indicated that the parameter space was too large for Eponine to be directly used efficiently in multi-component protein motif discovery (data not shown), which could be explained by the fact that amino acid alphabet is 5 fold larger than the DNA one, having high noise levels to be analysed with this tool. In most of these experiments, the system never converged automatically, and it contained "negative" motifs (data not shown).

In order to limit the problem size, I have come up with a hybrid, semi-guided, two-step procedure involving the probabilistic motif discovery tool NestedMICA, as well as the Eponine tool which can build multi-component hierarchical motif models to describe complicated sequence structures with its machine learning strategies. In the first step, I use NestedMICA to find some monopartite motifs, then by expanding those sequence regions by around 20 amino acids from both sides, where there is a significant match of a reported NestedMICA motif, I construct a new dataset composed of sequence chunks that have an instance of the used single-part motif. In the second phase, Eponine is run on this filtered dataset containing the positive samples, and also a negative dataset which has the same number of samples but not containing any motif hit.

To preserve the general sequence characteristics of the positive set across the negative set, sequence samples in the negative set are obtained from the same protein dataset so as to prevent Eponine from finding motifs that could possibly be reflecting potential compositional differences of the two sets.

Because I have sequence chunks with fixed lengths, each having a monopartite motif at the middle, Eponine was run in the EAS mode. After all, the aim is to find a multi-component motif model based on a reported NestedMICA motif whose position is known. The anchor point was specified as the maximum scoring point when scanned with the monopartite NestedMICA motif.

3.2.4 Using amino acid composition

It is possible to group amino acids according to their physical and chemical characteristics. If there are similar amino acids, one question to ask is whether grouping similar amino acid residues together, and then calculating the composition of the 'labels' of these groups rather than finding occurrence rates for each of the 20 amino acids could be a better approach or not. This brings two complications: determining the optimal number of such groups, and deciding which amino acid letters will be classified under which group. I used three amino acid groupings suggested by Thomas & Dill (1996), found by an iterative procedure involving "energy" scores calculated by iteration until they correctly discriminate a set of known protein folds from decoy conformations. Table 3.3 shows two types of amino acid groupings from Thomas & Dill (1996) and one grouping I formed based on general amino acid characteristics.

In the SVM, I kept all other features, except that the composition values

	Group 1	Group 2	Group 3
1	VILMF	VILMFWYA	ILVM
2	HQN	GPSTHQN	TSNQ
3	C	C	EDKRH
4	ED	ED	WFYP
5	RK	RK	C
6	A		GA
7	G		
8	WY		
9	P		
10	ST		

Table 3.3: **Alternative amino acid groupings used in composition calculation.** Groups 1 & 2 are from Thomas & Dill (1996), while Group 3 was constructed based on general amino acid properties.

were calculated according to these amino acid groups rather than using the 20 amino acids directly. The performances of the SVMs in the experiments were evaluated by using 5-fold cross validation. All parameters of the kernel function were optimised for each type of amino acid grouping I used, as in the optimisation of the actual SVM I used (see the section below, 3.2.6).

It turned out that grouping amino acids according to their physical and chemical properties is not particularly helpful (see page 95 in the Results section), so instead, 20 values have been computed to demonstrate amino acid composition statistics for each sequence.

3.2.5 Using transmembrane topology predictions

Apart from using amino acid composition and bit scores of motifs discovered by NestedMICA and Eponine, predicted transmembrane topology statistics were used as well to create Support Vector Machine feature vectors. Transmembrane

region predictions were reported by the 2c version of the TMHMM transmembrane topology prediction program (Krogh *et al.*, 2001). TMHMM was run in the "short statistics" mode. Amongst the reported TMHMM statistics, I included the following reported features:

- the number of predicted transmembrane helices
- the expected number of amino acids lying in transmembrane helices, considering the entire sequence
- the expected number of amino acids lying in transmembrane helices, considering only the first 60 N-terminal amino acids

Before using these reported numbers in the SVM, they were normalised with respect to the length of the input sequence considered.

3.2.6 Training and testing of SVM

I used a popular open source implementation of SVM, *libsvm* (Chang & Lin, 2001), in the multi-class predictor Lokum. In the parameter optimisations carried out to maximise the performance of each tried SVM application, *libsvm* performed slightly better than the other popular SVM applications I tried, namely, SVM^{light} (Joachims, 1999) and BSVM (Hsu & Lin, 2002).

Eventually, a radial basis kernel function (RBF) was used in *libsvm* after a systematic evaluation of a selection of kernel functions. In a similar way, I performed a grid search to optimise the gamma (g) and cost (C) parameters of this kernel function (Figure 3.3). The training and performance assessment of the SVM involved a 5-fold cross validation procedure in which the data were

divided into 5 portions; 4/5 of which were used for training and 1/5 for testing, using a particular portion for testing at a time in each of the 5 cycles. All protein scores coming from different features have been normalised to have a minimum value of -1 and a maximum value of 1, before the SVM software was run. The individual SVMs constructed to give an idea about the contributions of motif scores, composition and structural information were trained with 4/5 of the data. Kernel parameters of each SVM using a particular type of feature has been optimised, too, before I tested the SVMs with the remaining 1/5 portion. During kernel parameter optimisation, 3-fold cross validation was used for faster analysis.

3.2.7 Evaluation of Lokum predictions

The reported overall accuracy is the arithmetic mean of the correctly classified sequence percentage in each cross validation iteration. Sensitivity (SN), specificity (SP) and Matthew's Correlation Coefficient (MCC) (Matthews, 1975) values were calculated for each predicted class according to the formulae given in Equations 2.3, 2.4 and 2.5, respectively.

3.3 Results

By using NestedMICA, I found many motifs from different localisation datasets (see Appendix A for sequence logos of these motifs). Not all of these motifs ended up being used in Lokum, however: discovered motifs were assessed for their discriminative powers (see Section 3.2.2), and those not contributing to localisation prediction were filtered out. Figure 3.4 shows some mono-partite localisation

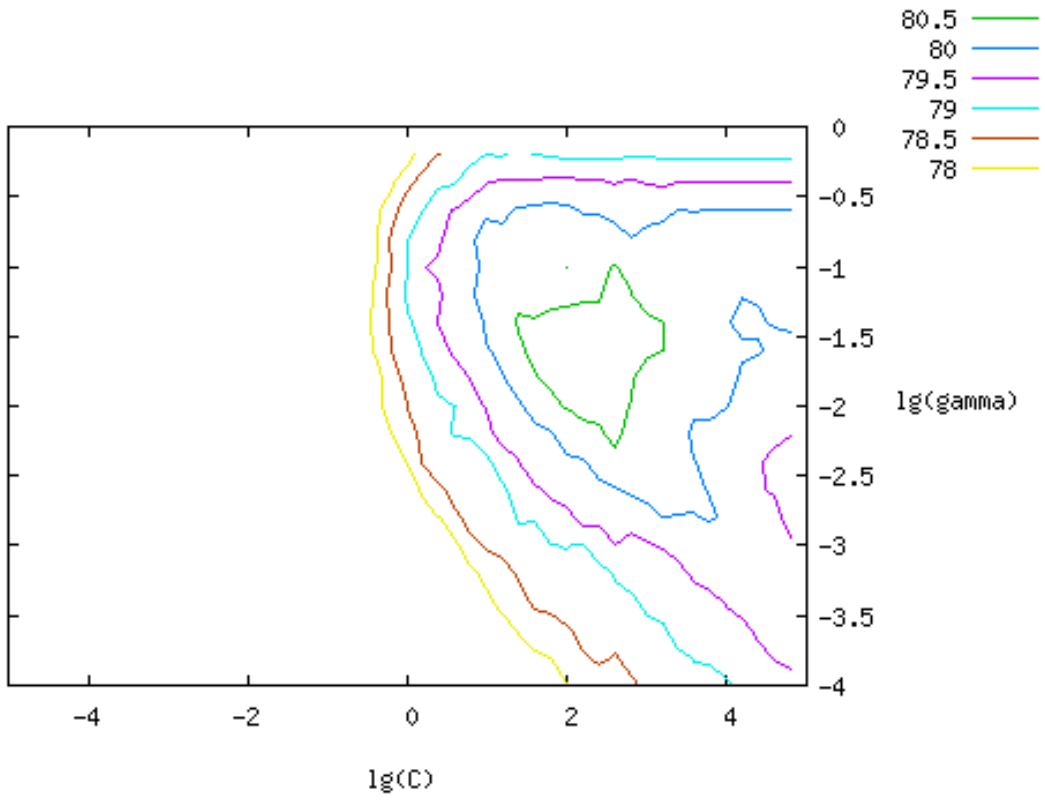


Figure 3.3: **SVM kernel parameter optimisation.** The plot shows an example set of percent accuracy contours formed by different values of the gamma (g) and cost (C) parameters (given in $\log_2$) during the optimisation of a radial-based kernel function (RBF) used in the SVM. Different pairs of g and C may produce similar percent accuracy rates, hence the countours. The specific example shown is for the animal protein version of Lokum. Accuracies plotted have been rounded to the nearest lower half values, i.e., an accuracy of 80.78% was considered in the 80.5% group of accuracies for plotting. Increasing the number of cross-validation iterations can increase the perceived performance (see text for the actual percent accuracies attained for different organisms using cross validation).

signals I used in Lokum. These motifs, represented as sequence logos here, correspond to some known localisation signals which are mostly characterised as regular expressions in literature. Other longer, and probably mostly unannotated, part-of-domain motifs that were used in Lokum can be found in Figure 3.5. Lokum prediction server is available online for public use at:

`http://www.sanger.ac.uk/Software/analysis/lokum/`

The discovered localisation related motifs that were used by Lokum can be downloaded in NestedMICA's XML format (.xms) from the same web page. See Section 3.5 for more information on the Lokum web server.

3.3.1 Discovered monopartite motifs

As plasma membranes have a highly hydrophobic region within their transmembrane helices (Figure 3.4c), which is very similar to hydrophobic regions of signal peptide sequences (Figure 3.4b), only the latter was used in the predictor. The signal peptide (SP) that is found in most of the secretory pathway proteins can be thought of consisting three parts: an N-terminal part (n-region) which can vary in length and has a net positive charge, a central hydrophobic core (Figure 3.4b), and a c-region which features a "-3 -1" rule (von Heijne, 1986) indicating the conserved positions with respect to the cleavage site (Figure 3.4a).

Figure 3.4k shows a good example of how NestedMICA can be efficiently used in short functional protein site finding. The depicted 4-position PWM looks quite similar to the cleavage site of a previously reported long chloroplast transit peptide (cTP) sequence logo (Figure 3.6) which was obtained by aligning the N-terminal regions of 62 chloroplast sequences with known cleavage site positions

Name	Motif
a) SP cleavage site	
b) Hydrophobic part of SP	
c) Transmembrane helix hydrophobic core	
d) N-linked glycosylation (1)	
e) N-linked glycosylation (2)	
f) ER retention	
g) C-terminal signal for recycling into ER	
h) Nuclear signals	
i) Nuclear signals	
j) Peroxisomal targeting signal 1 (PTS1)	
k) Chloroplast transit peptide (cTP) cleavage site	

Figure 3.4: **Some of the protein localisation related signals as recovered by NestedMICA.** Each motif has a maximum information content of 4.3 bits per position. Amino acids are drawn in four colours: hydrophobic residues are depicted in orange, hydrophilic and polar ones in green, acidic ones in pink, and finally basic amino acids are in blue.

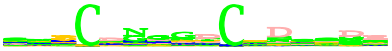
#	Dataset	Motif	Location scanned
1	mitochondrial		
2	plasma membrane		
3	plasme membrane		
4	lysosome		
5	lysosome		
6	golgi		
7	peroxisomal		
8	nuclear		
9	vacuolar		
10	vacuolar		
11	vacuolar		
12	vacuolar		
13	vacuolar		
14	vacuolar		
15	vacuolar		
16	chloroplast		
17	chloroplast		
18	chloroplast		

Figure 3.5: **Some of the unannotated signals, or part-of-domain motifs reported by NestedMICA.** These motifs were discovered from localisation datasets given on the second column. Sometimes the motifs were scanned in certain positions on protein sequences, rather than using the whole sequence (last column).

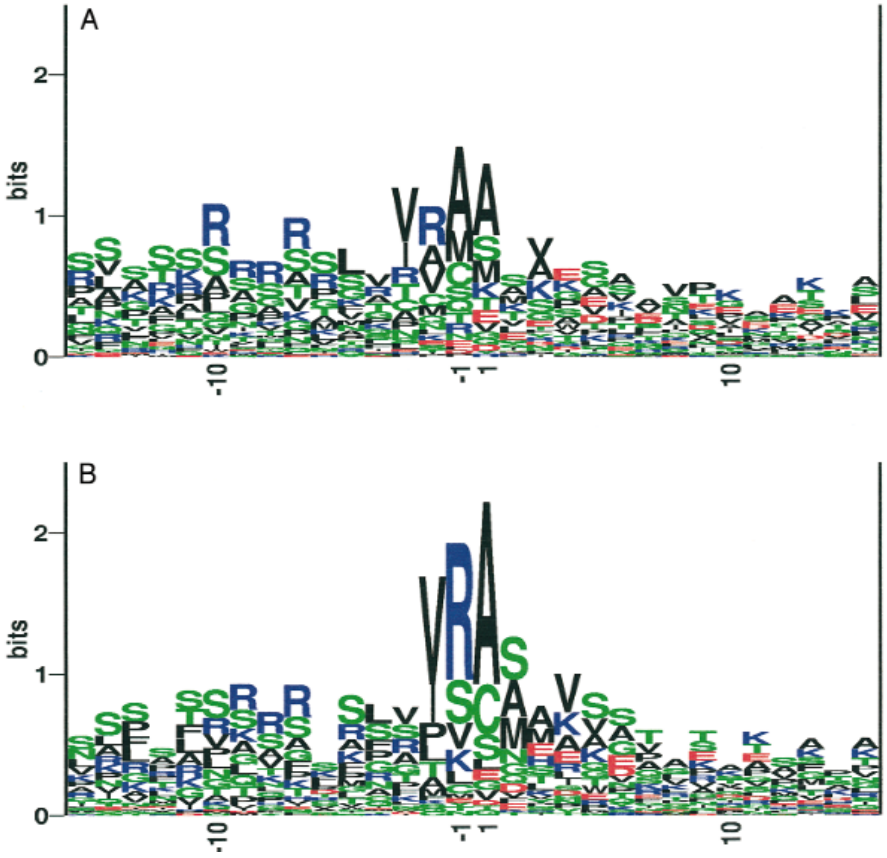


Figure 3.6: **Manually constructed motifs that are used in the ChloroP predictor.** This figure has been reproduced from the chloroP paper by Emanuelsson *et al.* (1999). The sequence logos were constructed from the 62 sequences used in the cleavage site predictor (chloroP) development. The sequences are aligned around their SWISS-PROT annotated cleavage site (top logo) and around the predicted cleavage site (bottom logo). Note the similarity between the motif shown in Figure 3.4k which is discovered automatically by NestedMICA and the conserved cleavage region of the manually aligned ChloroP logo in this figure (bottom).

that were kept fixed in the alignments (Emanuelsson *et al.*, 1999).

3.3.1.1 Contribution of N-linked glycosylation signal

Investigating the 3-letter motifs reported (Figure 3.4d-e), I found that these motifs correspond to the N-linked glycosylation signal which is found in two forms:

there is an Asparagine (N) residue in the first position followed by a non-conserved position, while the third position, determining the sub-variant, is occupied by either a Threonine (T) or a Serine (S) residue. Given that this is only a 3-letter motif, the chances are a fraction of its contribution to predictions could be due to some compositional effects. Namely, sequences having more number of the amino acid letters N, T or S, for example, could get higher scores when scanned with this motif, although, in reality they may not feature a glycosylation site. To investigate if there is a significant contribution coming from this motif apart from its compositional effects, I built artificial 3-letter motifs by inverting the positions of residues in this motif. Figure 3.7 shows the ROC curves measuring the classification power of the N-linked glycosylation motif, along with the shuffled motifs which of course retain the same composition as the original. The unshuffled original motif showed a better performance than all the other 5 possible variants, which indicates that using the N-linked glycosylation motifs is useful in computational protein localisation predictions, although it may not be directly involved in protein sorting processes as previous studies have demonstrated (see introduction).

3.3.1.2 Alternative ER retrieval

When NestedMICA was run on a dataset containing C-terminal ER sequences of length 20 aa, it reported the [KH]DEL motif shown in Figure 3.4f. When it was asked to find two motifs from the same region, instead of reporting a different or a weak motif (see the discussion on "null motifs" in Section 2.3.5), it reported another motif that looks like the first one, with the first residue being quite weak.

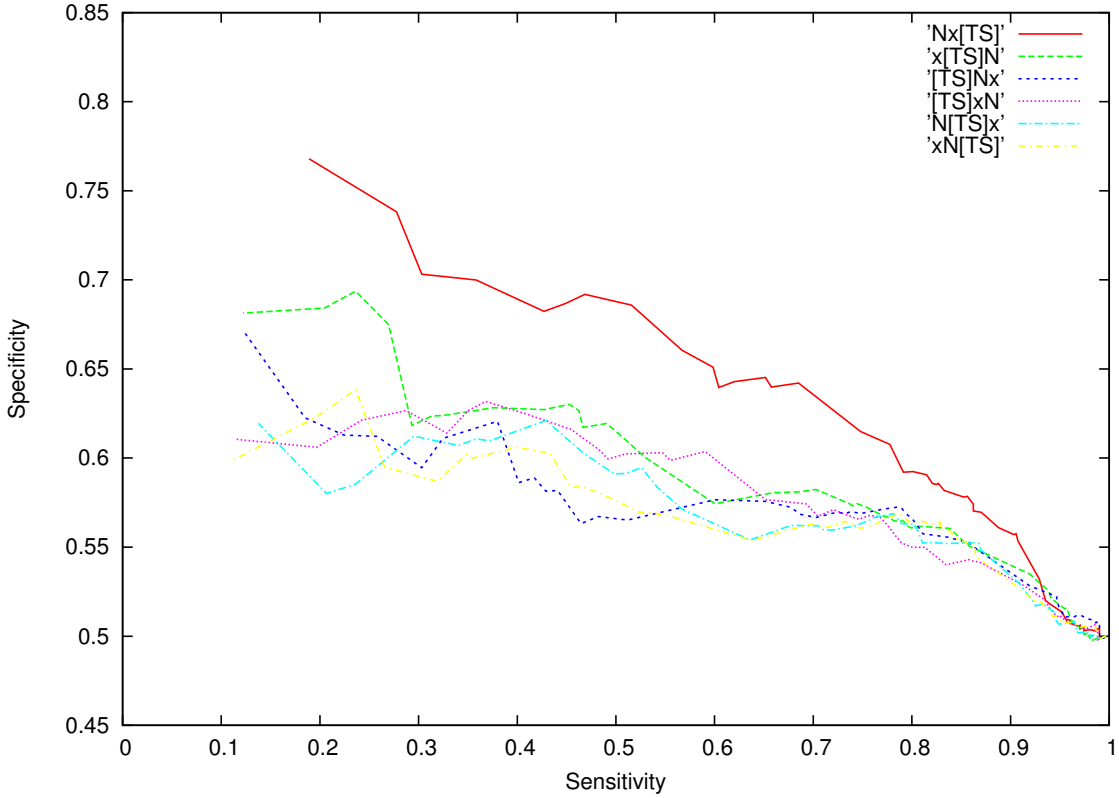


Figure 3.7: **ROC plots showing the contribution of the N-linked glycosylation motif in binary classification between nonredundant 509 plasma membrane and 509 mitochondrial protein sequences taken from the Multiloc datasets.** The curves correspond to the sensitivity (x-axis) and specificity (y-axis) values of multiple classifications performed by using a range of threshold scores. Each sequence was scored according to the best hit of the glycosylation motif and also the best hits of each of the derived PWMs obtained by shuffling the original motif's positions. This way I can evaluate a signal's performance with respect to the contribution of composition which is conserved in all the derived motifs. The red solid line shows the ROC for the original motif, while the dashed lines represent the shuffled PWMs' ROCs. Each motif's consensus sequence is shown in the legend, where the [TS] notation means there is either a T or S at that position, while 'x' indicates an unconserved position.

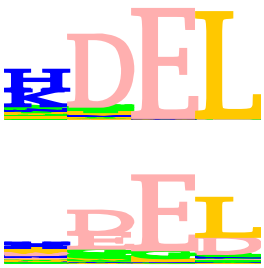


Figure 3.8: **The two C-terminal ER retention motifs reported.** The two motifs differed in their first residues, which may indicate that while some ER sequences have either K or H at position 1 of their C-terminal ER retention signals, some of them simply have not conserved the first position of this signal.

Figure 3.8 shows both motifs reported in this second run. This may suggest that while some sequences have either K or H at position 1 of this signal, in the others there is no preferred amino acid residue for this position, and that for them this signal is practically three amino acid longs.

NestedMICA has a useful feature which enables the user to find motifs other than a set of user-supplied motifs that are ignored during the program's motif search if they are found in the input sequence. When I run NestedMICA on a set of 20 amino acid long C-terminal ER amino acid sequence chunks by masking the [KH]DEL PWM found before, I came across an Arginine (R) and Lysine (K) rich motif that is shown in Figure 3.4g. While investigating some possible explanations to this motif in the literature, I found that Pelham (1995) had previously demonstrated that [KH]DEL is not the only C-terminus signal ER proteins might possess: a similar mechanism recycles escaped ER membrane proteins that have a loosely defined lysine (K)-rich, 4 amino acid long signal.

This reported NestedMICA PWM which does not have a clear consensus sequence could possibly be linked with this second ER retrieval mechanism. Including this motif in the SVM had a slight contribution ($< 1\%$) in the overall prediction performance.

3.3.1.3 Scanning motifs in certain positions

Some localisation signals have specific positions in sequences. The ER retention signal (Figure 3.4f), for example, is located at the far C-terminal end. Therefore, while scanning and scoring sequences for the presence of such motifs, only specific regions have been considered. In the case of the ER retention signal, this was the last four residues on the C-terminus. The SP cleavage motif (Figure 3.4a) was scanned in a window of 50 N-terminal amino acid positions. Similarly, the hydrophobic-residue rich motif of Figure 3.4b has been scored only within the 20 N-terminal sequences. Scanning PWMs in specific sequence regions where they are more likely to be present has a significant advantage over scanning them in the entire sequences. Figure 3.9 demonstrates one such example of how well motif b of Figure 3.4 can discriminate between redundancy reduced 841 extracellular and 841 cytoplasmic proteins, where two ROC curves are plotted using scores obtained by scanning the motif in whole-length sequences, and only in the first (N-terminal) 40 amino acid region, respectively.

3.3.1.4 Scoring multiple instances of motifs

In constructing the SVM vectors, in addition to using the maximum motif score corresponding to the sequence position where the best match occurs, I used the second best scores for the core NLSs (Figure 3.4h-i), and also for the N-linked

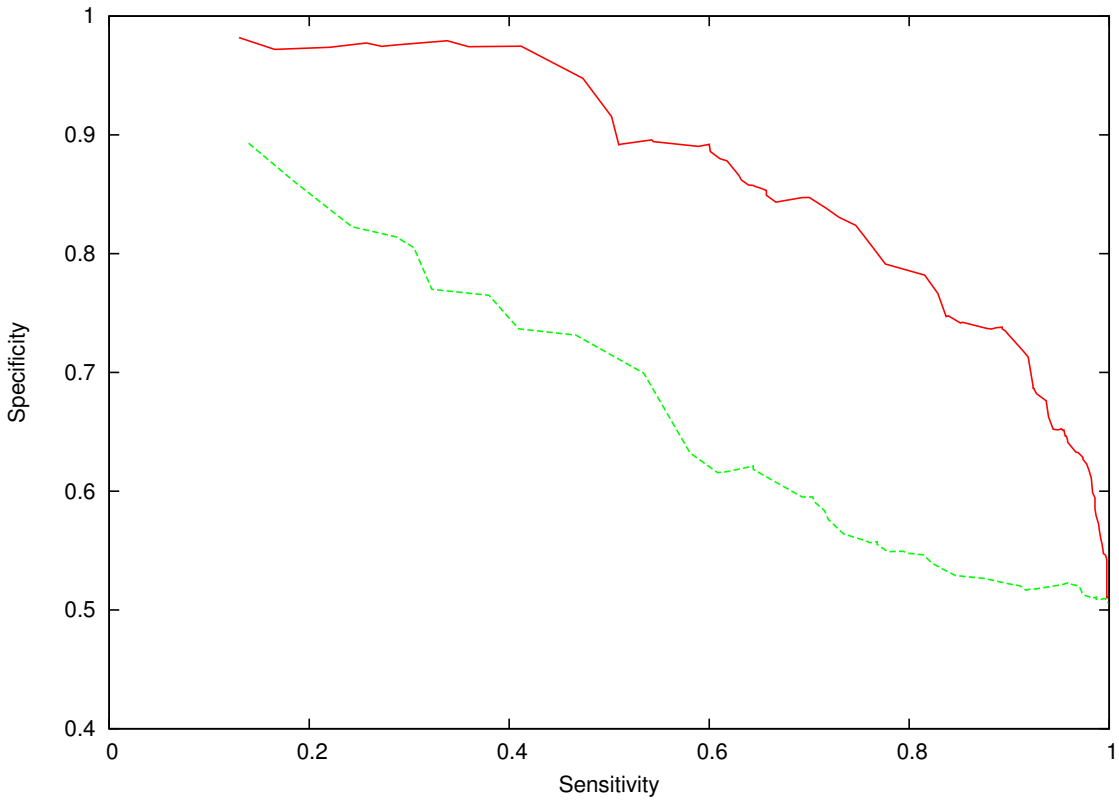


Figure 3.9: **A ROC curve showing the effect of scanning sequences with PWMs in certain segments only.** The plot shows different sensitivity (x-axis) and specificity (y-axis) values obtained for a range of score thresholds, indicating how well extracellular proteins can be discriminated from cytoplasmic proteins by using motif b of Figure 3.4. The red line is obtained when sequences were scored using only chunks of 40 N-terminal amino acids, while the green line represents the reduced performance attained when full-length sequences were scanned to obtain the maximum score.

glycosylation motifs (3.4 d-e). With this addition, I observed a significant increase in the overall classification accuracy, suggesting that some of the identified signals can possibly exist in more than a single region across a sequence. Other motifs did not even slighly increase the overall accuracy when I additionally used their second best scores.

3.3.2 Bipartite motif models

3.3.2.1 Bipartite NLS

As described in the methods section, I used a semi-guided procedure where I used both Eponine and NestedMICA to characterise such motifs. Figure 3.10 shows one possible model to describe a bipartite NLS signal. Generally, individual motif components do not have to have fixed positions in Eponine models; instead in Eponine's EAS models they are attributed with positional distributions with respect to an anchor point as described in Section 3.2.3. These Gaussian distributions reflect a motif's occurrence frequency within an optimal sequence range. The variations in the distributions shown on Figure 3.10 are quite minimal, indicating that relative sub-motif positions in this particular NLS model usually vary at most by a couple of residues.

In "nuclear versus others" type binary predictions made to assess the contribution of individual nuclear motifs, this bipartite NLS motif by its own classified correctly 141 nuclear sequences that the other mono-partite motifs shown in Figure 3.4h-i could not predict alone. Raw motif score thresholds used in these two-way classifications were chosen such that they maximise the corresponding MCC values computed to measure correct classification rate. The "others" sequence

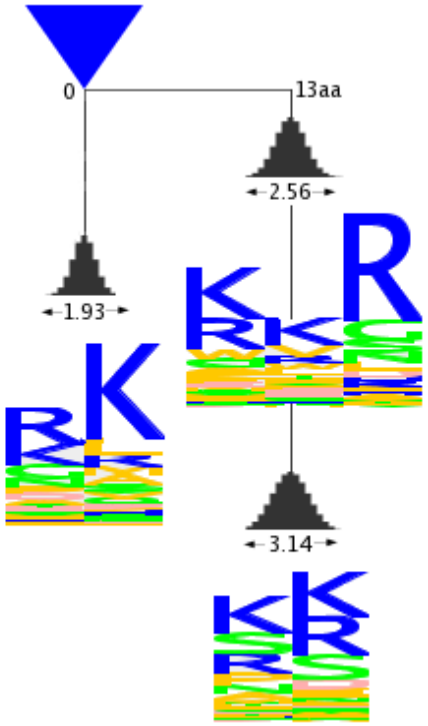


Figure 3.10: **Schematic representation of the Eponine bipartite NLS model.** The constraint distributions, the sequence logos and the relative positions of the individual components of the model are shown with respect to an anchor point (blue triangle). The central parts of the two main branches in the model are separated by 13aa's as shown. The model tolerates each sub-PWM to change position within the depicted probabilistic distribution width.

set in this particular experiment was compiled from the remaining 8 localisation datasets contributing in roughly equal numbers, and contained the same number of sequences in the tested nuclear set, 836.

3.3.2.2 Bipartite PTS2

To find a bipartite Eponine model for PTS2 (see Section 3.1.1), I followed the same procedure in modeling the bipartite NLS. However, it was more difficult for NestedMICA to discover the individual components of this weak bipartite motif,

each consisting of only a couple of adjacent residues, to enable me to perform the sequence filtering step in the multi-component model finding methodology (see methods) before running Eponine.

To investigate why NestedMICA failed to identify this motif or its components, I scanned the PTS2 regular expression "[RK][LVI].....[HQ][LA]" (Section 3.1.1) in 157 peroxisomal sequences I used. PTS2 is normally regarded as an N-terminal signal, but surprisingly I could locate only 4 hits within the first 50 amino acid N-terminal regions of these sequences. There were only a total of 31 matches of this regular expression when it was scanned in the whole-length sequences. This low abundance rate could explain why this weak motif, having two not well conserved amino acids on either side separated by 5 "spacers", could not be found by NestedMICA.

As an alternative, I ran Eponine on a dataset consisting of amino acid chunks matching the regular expression [RK][LVI].....[HQ][LA] of this motif. However, neither plotting ROCs to assess the obtained model's discriminative power from other types of proteins, nor the principle component analyses (see methods) I have performed suggested any performance gain from using this model. This indicates that this particular less conserved motif could be found in other classes of proteins by chance, and therefore it is not disjunctive enough in localisation prediction.

On the other hand, although the C-terminal PTS1 motif (having the short but conserved "SKL" form) that is shown in Figure 3.4j was not present in the majority of peroxisomal proteins, whenever a motif hit was found in the far C-terminal region, its selectivity was high, namely, it was most of the time capable

of discriminating a peroxisomal protein from another type.

3.3.3 Golgi N-terminal transmembrane topology prediction statistics help in localisation prediction

Knowing the transmembrane topology of a protein contributes to its localisation determination, since most of the cytoplasmic proteins will not contain as many membrane-spanning regions as plasma membrane proteins, for example. I found that even for different proteins of the secretory pathway where transmembrane regions are abundant, this could be used as a distinguishing feature.

Golgi does not have an apparent targeting or retention signal, but I observed that TMHMM, which may not distinguish between a signal peptide (SP) and a transmembrane (TM) helix, predicted at least one TM helix for 91% of the sequences in the Golgi dataset, 86% of which were predicted to be crossing the membrane once, while only approximately half of the ER sequences had at least one predicted TM helix. An overwhelming majority (97%) of plasma membrane sequences were predicted to possess at least one TM helix, too, but these were distributed across the sequence unlike in the Golgi sequences. Figure 3.11 shows the expected number of amino acids among the first 60 N-terminal residues that fall within a transmembrane region as reported by TMHMM for different protein classes. We know that these N-terminal transmembrane domain predictions are most likely signal peptide (SP) sequences responsible for targeting the majority of secretory pathway proteins into the ER after their synthesis. Unlike the other types of secretory pathway proteins, most of the Golgi proteins have their predicted membrane-spanning regions containing between 15 and 25 amino acids,

with a strong length preference of around 20 amino acid residues. This observation is justified by a previous study that showed that changing the length of the transmembrane domain of Golgi or plasma membrane proteins affected their protein localisation (Munro, 1995). In short, when incorporated into the SVM as described in Section 3.2.4 structural properties, such as the number and length of predicted TM structures in the N-termini and as well as in full-length sequences, clearly help Lokum in identifying protein localisation.

3.3.4 Effect of amino acid composition

In this work, in addition to using other protein features I use amino acid composition, too. However, this is not associated with the intention of by-passing possible annotation errors with this choice; instead, it is mostly to make advantage of the biological fact that proteins in a certain compartment can possess similar macroscopic properties such as composition, possibly for better interacting with their environment. As mentioned in the introduction section of this chapter, many previous studies have used amino acid composition as a strong sequence-level attribute that can be used as a distinguishing feature in subcellular localisation prediction. I used normalised amino acid frequencies to convey this macro-molecular characteristics that would presumably be similar in proteins sharing a common compartment. Proteins in different localisations can bear different predilections for certain amino acid residues, as the plots in Appendix B demonstrate.

Using the first type of amino acid grouping suggested by Thomas & Dill (1996) (see page 77 in the Methods for more detail), instead of the 20 amino acid letters

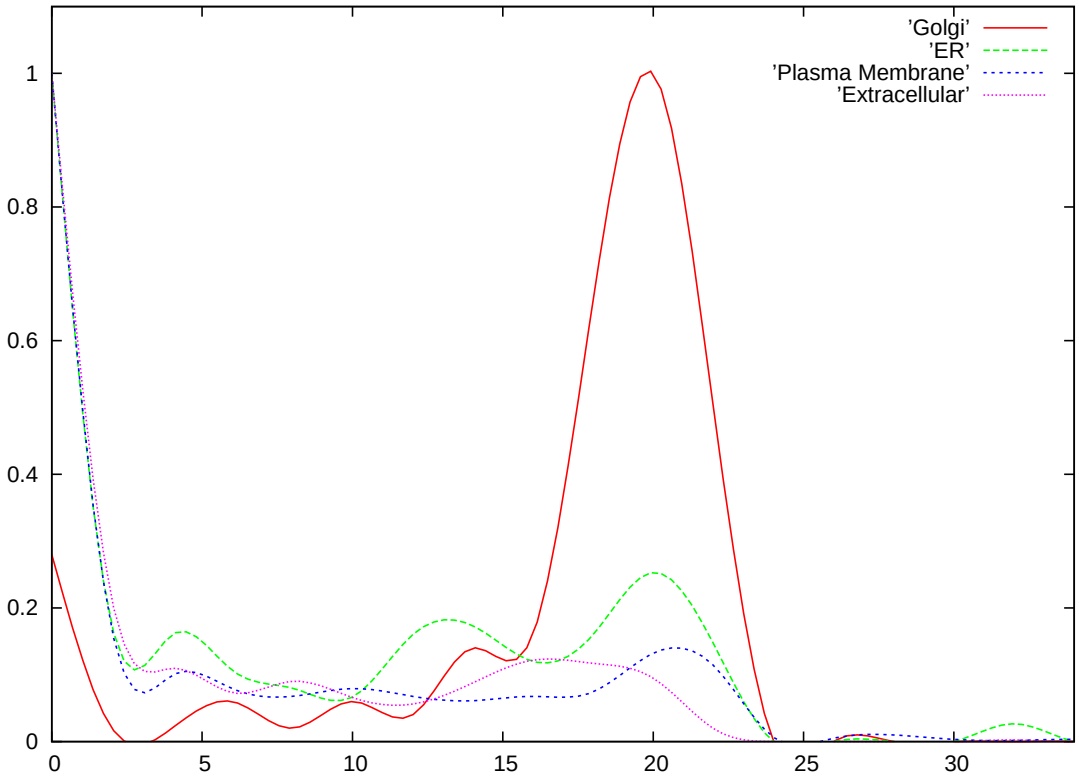


Figure 3.11: **The expected number of amino acids in the first 60 N-terminal residues to be part of a transmembrane region.** Plotted using a bin size of 5, the normalised histogram shows the predicted number of amino acids in the first 60 N-terminal residues lying within a transmembrane region as reported by TMHMM. Distributions for different types of secretory pathway proteins are shown. Plasma membrane proteins generally have a larger number of membrane-spanning regions spread across their entire amino acid sequence (see text), whereas Golgi sequences tended to have a single predicted TM helix in their N-termini, demonstrating a strong total length preference of around 20 amino acids.

in the calculation of composition, I was able to get a maximum correct prediction rate of 77.61% from the SVM using 5-fold cross validation, which is about 4% less than the result obtained from using all the amino acids in classifying the animal proteins. In addition to using composition, I also kept the other features like motif scores and transmembrane topology that I use in the general classifier. This group has 10 classes of amino acids (Table 3.3). Similarly, when I used the second type of amino acid grouping (3.3) from Thomas & Dill, where amino acids are categorised in 5 subgroups, the mean of the accuracy in the 5-fold cross validation tests was 76.54%. Finally, using the third type of grouping, also shown in Table 3.3, in which I mapped the 20 amino acids into 6 classes based on general physical and chemical properties of amino acids, I obtained an average correct prediction percentage of 75.81, in predicting the 9 animal protein localisation sets. These obtained figures are about 3-4% smaller than what I obtained by calculating the composition of each of the 20 amino acids without any grouping.

3.3.5 Lokum's performance

Table 3.4 summarises the performance of Lokum, in terms of the program's classification sensitivity (SN) and specificity (SP). Also, Matthew's Correlation Coefficient (MCC) values are given for Lokum, MultiLoc and PSORT in Table 3.5. SN, SP and MCC values were computed as explained in Section 3.2.7 on page 80. Individual cross validation sets used in the MultiLoc study were not available to enable me to perform a direct comparison. However, since Lokum is trained and evaluated using the same datasets of MultiLoc (Höglund *et al.*, 2006), for comparison I reproduced MCCs in the table for both MultiLoc and PSORT from

the MultiLoc paper where the latter programs are compared. For the SN and SP values of MultiLoc and PSORT please refer to the same article by Höglund *et al.* (unfortunately, this paper does not mention SN, SP and MCC values for all plant localisations).

3.3.6 Contributions of different features

To better understand the individual contributions of using motifs, composition and structural information, I stratified the prediction system by using only a particular type of feature at a time. I counted the number of correctly predicted protein sequences by running 3 different SVM predictors that use only motifs, only amino acid composition, and finally only transmembrane structure information. Figure 3.12 shows the proteins that were independently classified correctly by a single predictor, by any two, or by three of them. The Venn diagram tells us that about a third of the correct predictions can be achieved by either using only motifs or by composition alone. This indicates that the amino acid composition can be thought of as partially representing some of the motif information and vice versa. 13.7% of the predicted proteins can be said to be the easiest to predict, because they could be classified by any of the SVMs. More than a quarter of the proteins were predicted successfully only by the SVM using motif scores. The SVM that was trained only with structural information had the least number of correct predictions (3.8%) that the other predictors could not correctly classify.

Version	Localisation	Total sequence	Lokum performance		
			SN	SP	MCC
Animal	plasma membrane	1238	0.85	0.95	0.86
	mitochondrial	510	0.76	0.73	0.71
	nuclear	837	0.79	0.74	0.72
	cytoplasmic	1411	0.75	0.82	0.70
	ER	198	0.79	0.68	0.72
	extracellular	843	0.87	0.90	0.85
	Golgi	150	0.86	0.71	0.77
	lysosome	103	0.88	0.57	0.71
	peroxisomal	157	0.77	0.30	0.46
Fungal	plasma membrane	1238	0.86	0.95	0.86
	mitochondrial	510	0.75	0.72	0.70
	nuclear	837	0.77	0.75	0.70
	cytoplasmic	1411	0.75	0.81	0.70
	ER	198	0.82	0.68	0.73
	extracellular	843	0.85	0.90	0.85
	Golgi	150	0.84	0.71	0.77
	vacuolar	63	0.86	0.24	0.45
	peroxisomal	157	0.75	0.30	0.46
Plants	chloroplast	449	0.76	0.56	0.62
	cytoplasmic	1411	0.59	0.79	0.56
	plasma membrane	1238	0.86	0.95	0.86
	mitochondrial	510	0.69	0.66	0.63
	nuclear	837	0.75	0.73	0.69
	ER	198	0.81	0.67	0.73
	extracellular	843	0.84	0.88	0.83
	Golgi	150	0.83	0.72	0.77
	vacuolar	63	0.86	0.24	0.45
	peroxisomal	157	0.79	0.30	0.47

Table 3.4: **Prediction performance summary for Lokum.** Sensitivity (SN) and specificity (SP) and Matthew's Correlation Coefficient (MCC) values are given for Lokum. Lokum was trained and evaluated by 5-fold cross validation using the MultiLoc (Höglund *et al.*, 2006) datasets.

Version	Localisation	Lokum		MultiLoc		PSORT	
		MCC	Correct%	MCC	Correct%	MCC	Correct%
Animal	p. membrane	0.86	81.73	0.76	74.6	0.73	59.9
	mitochondrial	0.71		0.83		0.58	
	nuclear	0.72		0.73		0.54	
	cytoplasmic	0.70		0.68		0.43	
	ER	0.72		0.60		0.11	
	extracellular	0.85		0.77		0.72	
	Golgi	0.77		0.53		0.04	
	lysosome	0.71		0.48		0.18	
	peroxisomal	0.46		0.44		0.25	
Fungal	p. membrane	0.86	81.67	0.86	74.9	0.78	53.9
	mitochondrial	0.70		0.88		0.58	
	nuclear	0.70		0.73		0.54	
	cytoplasmic	0.70		0.69		0.43	
	ER	0.73		0.60		0.13	
	extracellular	0.85		0.73		0.68	
	Golgi	0.77		0.60		0.04	
	vacuolar	0.45		0.42		0.08	
	peroxisomal	0.46		0.43		0.25	
Plants	chloroplast	0.62	78.92	0.85	74.6	0.50	57.5
	cytoplasmic	0.56		0.70		0.42	
	p. membrane	0.86					
	mitochondrial	0.63					
	nuclear	0.69					
	ER	0.73					
	extracellular	0.83					
	Golgi	0.77					
	vacuolar	0.45					
	peroxisomal	0.47					

Table 3.5: **MCCs and correct prediction rates for Lokum, MultiLoc and PSORT.** The shown MCCs for MultiLoc and PSORT were taken from Table 3 of the MultiLoc article (data not available for all plant classes).

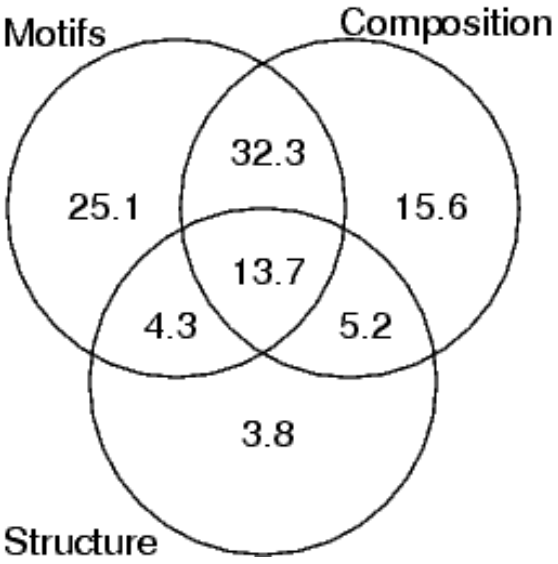


Figure 3.12: **Individual contributions of features used in the SVM.** This Venn diagram shows the percentage of proteins that could be correctly predicted by 3 individual SVM systems designed to use only motif scores, only amino acid composition, or only transmembrane statistics, respectively. The provided figures sum up to 100%, because only the distribution of proteins classified correctly at least by one predictor is given. The overlapping region between "Motifs" and "Composition" for example, indicates that amongst the proteins that could be predicted by at least one predictor, 32.3% of the "labeled" proteins could be successfully classified independently both by an SVM using only motif and another one using only composition information.

3.3.7 Contribution of disordered region predictions

Protein disorder regions are described and discussed in Chapter 4 (page 111) where I evaluated the use of disorder region statistics for use in sub-organelle localisation prediction. It has been suggested that inferring function improves when using patterns of native disorder in proteins (Lobley *et al.*, 2007). In order to assess the possible contribution of disorder prediction: i) I scanned the discovered localisation-related motifs in the predicted disorder regions to obtain a second set of motif scores, and ii) I considered the predicted disorder scores of sequence regions where a particular NestedMICA motif has a maximum score. Disorder region predictions were made using the RONN (Yang *et al.*, 2005) disorder prediction program (for the description of the software and methodology please see the dedicated Chapter 4).

However, adding these extra score sets (both at the same time or individually) to the SVM vectors resulted in no significant performance increase in the overall localisation prediction. After trying individual scores from both categories in different combinations, as performed by a systematic analysis, only a negligible maximum gain of around 0.01% could be achieved.

Using protein disorder predictions did not improve the overall prediction for the major localisation categories. This could be due to a number of reasons. Proteins can use different means to reach the same destination. Targeting into major cellular localisations can be achieved through general characteristics such as having a certain tendency in amino acid composition, which makes the disorder region statistics less effective for general localisation prediction. Nevertheless, as

shown in Chapter 4, knowing disorder regions could be useful for distinguishing proteins localised in different specific sub-organelle compartments.

3.4 Discussions

Computational prediction of protein localisation from amino acid sequence only is a challenging task not only because of some possible limitations in the methodologies, or even because of the lack of enough knowledge about the underlying biology. We know that proteins can migrate from a certain compartment to another, which does not permit a "one protein one localisation" correlation to always hold true. Besides, not all proteins have targeting signals, some are 'piggy-backed' and transported by other proteins which have the necessary signals (Wu *et al.*, 2000). Also, not all proteins from the same localisation categories show significant similarities in their general properties such as amino acid composition to enable one to make near-perfect predictions by only using these statistics. Therefore, one key factor in getting reasonable prediction accuracies lies in using as much relevant information as possible. When protein features such as localisation motifs, amino acid composition, or structural information etc. are used in combination, perhaps each bit would be characterising a certain number of protein classes better, but also their synergy would result in better overall prediction quality by possibly reducing some of the false positive predictions that individual feature components would otherwise produce.

Motifs like the N-linked glycosylation signal, one of the oldest known protein signals (Prosite id: PDOC00001), could be of great help in localisation prediction, even though they may not be directly involved in protein targeting. The N-

linked glycosylation process which normally takes place in the ER lumen aids us in predicting secretory pathway proteins when combined with the extra motifs found.

Representing motifs as PWMs rather than regular expressions is advantageous. As mentioned in the introduction, there are two major types of peroxisomal targeting signals (PTS). The first identified PTS is the C-terminal SKL-type signal. However, in some cases, it can take the form of a similar tripeptide, namely "KKL" (Takada *et al.*, 1990) and in some other eukaryotes it could be "SQL" (Purdue *et al.*, 1992), "NKL" (Lumb *et al.*, 1994; Oda *et al.*, 1987), or "SSL" (Motley *et al.*, 1995). Existence of many such possible variants clearly indicates that localisation motifs represented in regular expressions cannot be as efficient as probabilistic representations. PWMs, such as the PTS1 motif shown in Figure 3.4j, can potentially tolerate slightly differing forms by allowing a certain degree of sequence variation due to their probabilistic construction.

Although the *ab initio* Lokum does not use any database look ups to detect proteins matching a certain Prosite or NLSdb motif, its performance in assigning eukaryotic proteins into the correct localisation category was better for most of the localisation categories than the other multi-class predictors compared. I showed that by combining features including motifs represented as PWMs, amino acid composition and transmembrane topology statistics, one can get very reasonable (as high as 81%) prediction accuracies. As I demonstrated with the glycosylation motif example, protein motifs that are not directly involved in protein sorting could be used as secondary signals, too. In some cases, composition can substitute the information coming from a signal, but most of the time using direct biological

localisation signals along with composition and structure statistics proved to be more efficient.

By leaving out one sequence at a time and training a dedicated model by using the rest of the sequences to predict the localisation of that sequence, I was able to get an average correct prediction rate of 81.77% after repeating this procedure for each sequence in the entire dataset. This accuracy rate obtained by this "jack-knifing" methodology, however, only marginally differs from the reported correct prediction percentage of 81.73 (Table 3.5), which is obtained from the 5-cross validation tests done for the animals category. On the other hand, the overall performance was calculated to be 79.5%, 80.5%, and 81.1% when I used, 2, 3 and finally 4-fold cross validation, respectively. This indicates that using 5-cross validation was adequate and that there is no need to further increase the number of cross validation test sets.

3.5 Availability

The Lokum protein subcellular localisation predictor is available for public use through a web server which can be reached at:

`http://www.sanger.ac.uk/Software/analysis/lokum/`

It allows users to either paste some sequences into a text box or upload a file of protein sequences in fasta format. A screenshot of the server can be seen in Figure 3.13. Users can specify the Lokum prediction mode (animals, plants or fungi) that they want to use for their sequences.

I wrote the public Lokum predictor as a Java servlet. It runs on a "Resin" dynamic web server on a Linux cluster, but it has been also tested on different

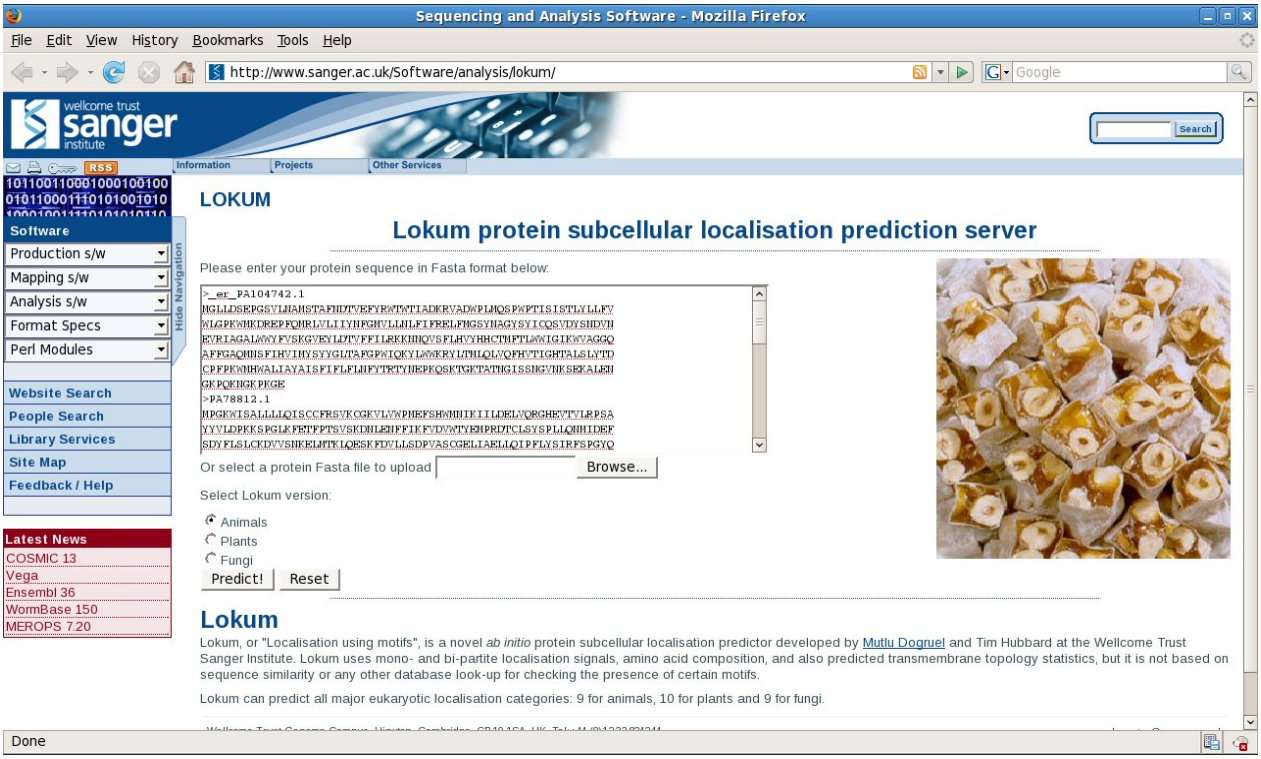


Figure 3.13: **Lokum prediction service hosted by the Wellcome Trust Sanger Institute.** Sequences must be uploaded either as a single fasta file, or entered into the text box in fasta format. Predictions are displayed in a separate page, following the submission of data.

platforms and using Tomcat, another popular web server. The servlet is based on the same command line version of Lokum, and also the same trained SVM classification model files. However, the prediction server works with a Java implementation of *libsvm* version 2.85, instead of the commonly used version written in the C programming language. No significant difference was observed between the predictions made by the two Lokum versions.

Interested users can download the protein motifs used in Lokum in Nested-MICA's XML format (XMS) from the Lokum home page.