Chapter 6

Conclusions

Motif discovery is an important step in protein functional annotation as it can help to identify different protein properties in curation of protein annotation. I adapted and extended NestedMICA for finding short protein signals, and compared its performance with the MEME tool. NestedMICA was tested on synthetic and biologically-authentic datasets produced by spiking instances of known motifs into a some random protein sequences. NestedMICA was also assessed at various conditions including using different input sequence lengths, target motif length, target motif number, and finally different motif abundance rates.

Generally NestedMICA recovered most of the short (3-9 amino acid long) test protein motifs spiked into a test set of sequences at different frequencies. All assessments experiments I performed showed that NestedMICA's motif discovery performance was better than MEME in terms of the number of correctly recovered motifs, although generally there was no significant difference in terms of the quality of recovered motifs by both of the compared programs. NestedMICA performed clearly well even in the discovery of relatively short motifs that exist in only a small fraction of sequences.

Protein subcellular localisation identification is another concrete key step in functional annotation. Most of the biologically inspired *ab initio* methods that have been developed to tackle this problem had either a limited number of localisation categories, or low prediction accuracies, particularly for eukaryotic sequences. Similarity-based prediction methods could be more reliable than *ab initio* predictors for sequences having annotated highly homologous counterparts in databases. However, predicting localisation for unseen, different proteins becomes a more challenging task for this type of prediction program. Furthermore, signal-based *ab initio* prediction efforts can give us more insight and clues about the underlying biology in protein targeting.

I developed a novel computational *ab initio* classification tool, Lokum, for protein subcellular localisation prediction, covering 9 major localisation classes for animal, 9 for fungal and 10 for plant sequences. It uses targeting and retention signal motifs reported by the probabilistic motif discovery tool NestedMICA, and other protein features including transmembrane topologies and amino acid composition. Lokum does not use sequence similarity, or any other *a priori* knowledge such as known nuclear localisation signals by searching databases. Additionally, we propose a multi-component, probabilistic model tolerating positional shifts for the bipartite nuclear localisation signals (NLS). To find the bipartite NLS, we added protein support to Eponine, a tool originally written for mammalian transcription start site modeling. We also show that using the N-linked glycosylation motif, which was amongst the motifs detected by NestedMICA, can contribute to localisation prediction.

Combining all these features in a Support Vector Machine (SVM), we get an

average correct prediction rate of more than 80% for nine animal, nine fungal and ten plant protein localisation classes in 5-fold cross-validated tests performed on an eukaryotic dataset. Finally, a web service has been implemented for public use.

In Chapter 3, I showed that including reported statistics from transmembrane prediction programs can increase prediction accuracy in automatic *ab initio* classification of protein subcellular localisation. A large number of transmembrane proteins follow the secretory pathway and end up in localisations such as ER, Golgi, plasma membrane or extracellular space. Plasma membrane proteins have a larger number of membrane-spanning regions than the other classes of proteins, as shown in the same chapter. Therefore, it is actually not surprising that transmembrane topology prediction can improve localisation.

Motifs reported by NestedMICA and Eponine have been more useful than any other component in the prediction system. In addition to the reported motifs that I could associate with known localisation signals, a couple of three-letter PWMs were discovered from a set of plasma membrane sequences, which turned out to be the two variants of the N-linked glycosylation site motifs. Some of the discovered motifs, such as these glycosylation motifs that are known not to be directly involved in localisation, also increased the prediction performance, because of their differing abundance rates in different types of proteins.

In Chapter 4, I showed that it is reasonably possible to predict more specific, sub-compartmental localisation categories, by showing that proteins that spend at least some time in nucleoli can be distinguished reasonably well from the remaining nuclear proteins. In addition to the features used in Lokum, I used

protein disorder region predictions. As summarised in Section 3.3.7, using disorder prediction did not contribute significantly to the prediction of the general localisation categories. But I demonstrated that disorder prediction can be a useful feature in discriminating between proteins targeted into different sub-nuclear compartments. In fact, sub-dividing the main localisation categories to further fine tune localisation prediction can be said to overlap with the field of *ab initio* protein function identification, where disorder prediction has been shown to work (Dunker *et al.*, 2000; Lobley *et al.*, 2007; Wright & Dyson, 1999). Interestingly, the results obtained in Chapter 4 suggested that a larger number of nuclear localisation signals exist in the disordered regions of nucleolar proteins as compared to the disordered regions of other proteins in the nucleus. It should be possible to further exploit this phenomenon in the prediction of proteins localised in other sub-compartments.

An interesting observation we can make from Chapters 3, 4, and 5 is that there is a general tendency in protein amino acid composition to contain Lysine (K) and Arginine (R) residues at larger proportions as we move from the extracellular space towards the cytoplasm, and finally into the nucleus and other subnuclear compartments. If we consider the amino acid contents of extracellular, cytoplasmic and nuclear proteins, amino acids K (Figure B.12) and R (Figure B.2) are least abundant in extracellular, followed by cytoplasmic and then nuclear proteins, in this order. In Section 5.2.1.1 of Chapter 5, we saw that membrane spanning regions towards the cytoplasm become richer in K and R content compared to their parts on the opposite side, towards the extracellular space (Tables 5.1, 5.2 and 5.3). Finally, in Chapter 4, where I analysed the differences between

nuclear and nucleolar protein sequences, it turned out that nuclear proteins, which are confined in the sub-nuclear compartment of nucleolus, tend to contain a larger number of K and R amino acid residues (Figure 4.7).

Finally, as demonstrated by its application on transmembrane topology prediction, the introduced alternative transition probability optimisation method that I developed (Chapter 5) is a promising approach for use in any prediction program that utilises HMMs, including the classical problems of gene finding, secondary structure prediction, and so on.