CHAPTER 4

EPILOGUE

In this thesis, two life science problems have been tackled using information-theoretic pattern recognition techniques. The first problem involves the identification of single nucleotide polymorphisms (SNPs) that are useful for population inference. In other words, these SNPs make up to panels of ancestry informative markers (AIMs). The problem is difficult because the number of discrete-valued attributes is large while the sample size is small. As a result, the problem can be treated as an attribute selection problem. A round robin symmetrical uncertainty ranking technique has been developed for the task. The technique is proven to be a vital part of the AIM extraction protocol. This subsequently leads to the identification of AIM panels that are smaller than those previously reported.

The second problem of interest is a thalassaemia classification problem. In contrast to the first problem, this problem involves continuous-valued attributes. The procedure for solving the problem hence begins with information-theoretic attribute discretisation (Fayyad and Irani, 1993). Then informative attributes are identified via a correlation-based feature selection technique (Hall and Holmes, 2003). Finally, the classification model is constructed using a C4.5 decision tree (Quinlan, 1993). The result indicates that significantly high classification accuracy can be achieved. Overall, the results from both problems suggest that information-theoretic pattern recognition techniques are highly efficient and are proven to be useful for problems in life science.