Summary

The oxygen binding proteins are widely present in animals and in plants; it's responsible for binding and transporting oxygen throughout the body where it is used in aerobic metabolic pathways. These proteins are reported to be present in some of prokaryotes, whereas well established in eukaryotic systems. These proteins are Hemoglobin, Hemerythrin, Erythrocruorin, Hemocyanin, Leghemoglobin and Myoglobin. Many Oxy-proteins exhibit different functions such as oxygen binding, electron transfer, and metabolism. In this studies, Oxygen binding proteins with various prediction approaches, such as amino, dipeptide composition and position specific scoring matrix and achieved the accuracy was 85.12% , sensitivity 94.49%, specificity 75.97% and MCC 0.80. In dipeptide composition method, we achieved the maximum accuracy 85.12%, and 94.49%, 75.97%, 0.80 of sensitivity, specificity and MCC respectively. The PSSM profile of oxy and non-oxy proteins were generated for SVM prediction. The maximum accuracy was 89.20%, sensitivity 97.33%, specificity 79.86%, and MCC 0.85. Also we achieved the maximum accuracy in sub-class of oxy-proteins erythrocruorin, hemocyanin, hemerythrin, hemoglobin, leghemoglobin and myoglobin in AC, DC and PSSM methods.

The recent upsurge in microbial genome data has revealed that hemoglobin-like proteins (HbL) may be widely distributed among bacteria, and that some organisms may carry more than one HbL encoding gene. However, discovery of HbL proteins has been limited to a small number of bacteria only. This study describes the prediction of HbL proteins and their domain classification using a machine learning approach. Support vector machine (SVM) models were developed for predicting bacterial HbLs based upon amino acid composition (AC), dipeptide composition (DC) or position specific scoring matrix profiles (PSSM). In addition, we introduced for the first time a new prediction method based on Max to Min residue profiles (MM). The performances of the different approaches were estimated using fivefold cross validation techniques. Prediction accuracy was further investigated through confusion matrix and receiver operating characteristic (ROC) curve analysis. Bac-Hbpred, a web tool for Bacterial HbL prediction and classification has been developed and is publicly accessible at http://mamsap.it.deakin.edu.au/bac_hbpred/home.html

A support vector machine (SVM) model has been developed for predicting bacterial Hbs using five fold cross validation technique. We tried two step predictions, in the first step, we predicted the bacterial Hbs to distinguish them from non-bacterial Hbs using amino acid, dipeptide composition, PSSM and MM profile methods by which we achieved the maximum accuracy of 86.14% with 0.82 MCC (Mathews correlation coefficient), 83.02% with MCC 0.78, 90.20% with 0.89 of MCC and 86.28% accuracy with 0.83 MCC respectively. In the second step prediction, we attempted to classify the bacterial Hbs classification and individual domain by their amino acid, dipeptide and PSSM profile, and achieved the maximum accuracy. The average accuracy in each case was 85.76%, 80.45%, 87.93%, 85.46% of AC, DC, PSSM and MM respectively.

The uncharacterized (un-annotated) proteins such as blood proteins and plasminogen activator are also important. In this prediction studies we have applied amino acid and dipeptide composition methods for investigating blood proteins. The result demonstrated that the method can differentiate blood-proteins from non blood-proteins with greater accuracy of 90.57% in 0.89 of MCC at a default cutoff score of 0. The dipeptide composition method was also tested and achieved 91.39% accuracy with 0.90 of the MCC. The classification of blood protein prediction also tried, the detailed results are shown in chapter 5 (Result and discussion). The uncharacterized (unannotated) proteins sequences were tried, which retrieved from SwissProt protein database. The blood proteins of all models were detected only 31 out of 806 uncharacterized sequences.

In plasminogen activators, Pg-activators are serine proteases that cleave the plasminogen to produce two chains of active plasmin by a single proteolytic cleavage of Arg560-Val562 peptide bond. Plasmin is responsible for the degradation of blood clots. Broadly, the Pg-activators are classified into two types based on its function, as direct and indirect. The identification of Pg-activators is very important in molecular recognition. Amino acid composition based prediction of Pg- activators (SK, SAK, UK and tPA) against non Pg-activators resulted in a maximum accuracy of 88.37% with 95.24%, 83.50%, 0.87 sensitivity, specificity and Mathew correlation coefficient (MCC) respectively. In DC the SVM based model was able to achieve a maximum accuracy of 84.32% to 97.01%, 75.31% sensitivity, specificity, and 0.83 MCC. In position specific score matrix (PSSM) profile based prediction models for Pg-activators were also developed and achieved a maximum accuracy of 87.61% with 95.77%, 81.81%

sensitivity, specificity and 0.86 of MCC. In order to improve the performance of SVM models, a hybrid prediction method combining amino acid composition (AC) and dipeptide composition (DC) was also attempted to solve the Pg-activator protein prediction problem. Using the hybrid approach, accuracy, sensitivity, specificity and MCC were 85.63%, 97.71%, 77.06%, and 0.85 respectively. Furthermore, we also have developed a web server, which predicts the Pg-activators and their classification (available online at http://mamsap.it.deakin.edu.au/plas_pred/home.html). A total of 207 new sequences previously unseen by the SVM algorithm were obtained. All these sequences were correctly predicted as positive by all pg-activator models (AC, DC, PSSM and Hybrid models)

Therefore, our prediction results show that Oxygen binding proteins and non-annotated (blood and Pg-activators) are predictable with a high accuracy from their primary sequence. Our prediction performance was also cross-checked by confusion matrix and ROC (Receiver operating characteristics) analysis. A web server to facilitate the prediction of used proteins from primary sequence data was implemented. Our experimental results show that our approaches are faster and achieve generally a good prediction performance.

Keeping the increasing gap between protein with known sequence and know function the major aim of this thesis was to develop method(s) that can be used to predict function of unknown proteins. In order to achieve this goal; different methods have been developed which could be reliably used for proteome annotation.

The overall objective of this work was to develop improved and novel prediction methods for identifying potential candidates for oxygen binding proteins and uncharacterized (non-annotated) protein family such as blood proteins and plasminogen activators. The computational approach can provide a good alternative to experimental analysis. On the basis of this analysis, rules can be derived to model these processes and develop methods.

In conclusion, we have developed methods which may help biologist directly or indirectly in assigning function to newly sequenced proteins. In addition, the proteomes of few important organisms have been annotated using the developed methods. Overall methods developed in this thesis will be useful for Bioinformaticains and biologist involved in functional annotation of proteins.